THE UNIVERSITY
of EDINBURGH

Construct truncation due to suboptimal person and item selection: consequences and potential solutions

Aja Louise Murray

PhD

University of Edinburgh

2015

Contents

	Page
Declaration of own work	**v**
Acknowledgements	**vi**
Note	**vii**
Abstract	**viii**
Chapter 1: Introduction	**1**
Chapter 2: Construct truncation due to the use of clinically ascertained samples in autism spectrum disorder research	**14**
Chapter 3: How construct truncation on achievement may have distorted our understanding of the relation between conscientiousness and ability	**41**
Chapter 4: A comparison of alternative phenotypic proxies in tests of gene-environment interactions under construct truncation	**61**
Chapter 5: Discussion	**92**

List of Tables and Figures

Tables

Table 1.1: Some examples of construct truncation due to person sampling

Table 1.2: Examples of construct truncation due to item selection

Table 2.1: Extents of attenuation of symptom inter-correlations under simultaneous selection

Table 2.2: Descriptive statistics for ASD, non-ASD and combined samples for the 5 AQ domains

Table 2.3: Correlation matrix of the 5 AQ domains in ASD versus combined sample

Table 3.1: Correlations between IQ and conscientiousness at different levels of selectivity for educational achievement in adolescent sample

Table 3.2: Correlations between IQ and conscientiousness at different levels of selectivity for occupational achievement in adult sample

Table 3.3: Application of Thorndike case III to adolescent sample correlations

Table 3.4: Application of Thorndike case III to parent sample correlations

Table 4.1: Parameter values for IRT model used in simulation

Table 4.2: Performance of sum score, transformed score and IRT score latent trait proxies under different population biometric model

Table 4.3: Descriptive statistics for Well-being, Aggression and Intellectual Interests phenotypes

Table 4.4: GxM model fits for Well-being phenotype

Table 4.5: Parameter estimates from best-fitting models for Well-being phenotype

Table 4.6: GxM model fits for Aggression phenotype

Table 4.7: Parameter estimates from best-fitting models for Aggression phenotype

Figures

Figure 1.1: Population and truncated sample distributions for a phenotype

Figure 1.2: Scatterplot for variables in population and truncated sample

Figure 2.1: Contour plots of RSB and Soc before and after selection

Figure 2.2: Marginal distributions of RSB and Soc before and after selection

Figure 2.3: Histograms of AQ subscale scores in combined sample

Figure 2.4: Scatterplot matrix of AQ subscale scores in combined sample

Figure 3.1: Conscientiousness-IQ associations in adolescent sample

Figure 3.2: Conscientiousness-IQ associations in parent sample

Figure 4.1: Histogram showing the distribution of the sum score derived from generating item level data according to Eq. 3 with parameters in Table 4.1

Figure 4.2: Histogram showing the distribution of the transformed sum score derived from generating item level data according to Eq. 3 with parameters in Table 4.1 and then applying a $\log_{10}$ transformation

Figure 4.3: Histogram showing the approximate distribution of factor scores derived from generating item level data according to Eq. 3 with parameters in Table 4.1 fitting a graded response model, and then obtaining factor scores based on this model

Declaration of own work

I, Aja Murray, confirm the following:

(a) that the thesis has been composed by me, and

(b) that the work is my own, and

(c) the work has not been submitted for any other degree or professional qualification, and

(d) that any included publications are my own work, except where indicated throughout the thesis and summarised and clearly identified on the declarations page of the thesis.

Signed… **…….**

Aja Louise Murray

Acknowledgements

I am grateful to the many individuals who contributed in some way or another to this thesis. I am first of all grateful to my study co-authors for their invaluable contributions: Wendy Johnson, Bob Krueger, Bill Iacono, Matt McGue, Tom Booth, Renate Kuenssberg, Karen McKenzie, Michael O'Donnell, and Dylan Molenaar. For their mentorship, I am grateful to Wendy Johnson, Ian Deary, and Dylan Molenaar. I am grateful to the many friends and colleagues who provided support and/or comic relief as needed, especially Caterina Constantinescu, Amanda Jubb and Ingrid Obsuth without whom the PhD years and my subsequent transition to the post-PhD world of academia would not have been half as enjoyable. Finally, I am grateful to the many others in Edinburgh, Amsterdam and Cambridge who have helped create the stimulating research environments that I have been lucky to be a part of over the last 8 years and who helped shape my thinking and development in immeasurable and invariably positive ways. Deserving of special mentions, I am grateful to my supervisor Wendy Johnson who continually encouraged me to raise my game; to my parents George Murray and Karen McKenzie who provided many (solicited and unsolicited) sense checks on my research along the way; and to my partner Tom Booth for his love and support over the last 3 years.

Note

A version of Chapter 2 was published as:

'Murray, A. L., McKenzie, K., Kuenssberg, R., & O'Donnell, M. (2014). Are We Under-Estimating the Association Between Autism Symptoms? The Importance of Considering Simultaneous Selection When Using Samples of Individuals Who Meet Diagnostic Criteria for an Autism Spectrum Disorder *Journal of Autism and Developmental Disorders, 44*, 2921-2930.'

A version of Chapter 3 was published as:

'Murray, A. L., Johnson, W., McGue, M., & Iacono, W. G. (2014). How are conscientiousness and cognitive ability related to one another? A re-examination of the intelligence compensation hypothesis. *Personality and Individual Differences, 70,* 17-22.'

A version of Chapter 4 is currently under review as:

'Murray, A. L., Molenaar, D., Johnson, W., & Krueger, R. F. Dependence of gene-by-environment interactions (GxE) on scaling: Comparing the use of sum scores, transformed sum scores and IRT scores for the phenotype in tests of GxE.'

Abstract

Construct truncation can be defined as the failure to capture variation along the entire continuum of a construct reliably. It can occur due to suboptimal person selection or due to suboptimal item selection. In this thesis, I used a series of simulation studies coupled with real data examples to characterise the consequences of construct truncation on the inferences made in empirical research. The analyses suggested that construct truncation has the potential to result in significant distortions of substantive conclusions. Based on these analyses I developed recommendations for anticipating the circumstances under which construct truncation is likely to be problematic, identifying it when it occurs, and mitigating its adverse effects on substantive conclusions drawn from affected data.

Chapter 1: Introduction

Empirical research in the social sciences typically involves drawing a sample of participants from some target population, measuring constructs of interest in that sample, and conducting statistical tests on the resulting data. For inferences to the target population to be valid, it is assumed that the sample is representative of the target population in all relevant respects, including that the constructs of interest as they exist in that target population have been successfully represented by the questionnaires, tests, or measures administered. In the current thesis, I discuss a specific and common way in which this assumption is violated and the consequences that this has for theoretical inferences made under these circumstances: I discuss the problem of construct truncation due to inadequate item or person sampling.

Defining construct truncation

For the purposes of this thesis I define construct truncation as under-representation of the more extreme levels of a construct. Construct truncation can occur due to inadequate person or item sampling, sometimes these are referred to as 'type 1' and 'type 2' sampling respectively (Revelle & Zinbarg, 2009). In the case of truncation due to person sampling, individuals with the highest and/or lowest levels of the construct of interest are under-represented in the sample. In the case of truncation due to item sampling, a questionnaire does not include items capable of providing reliable measures of the highest and/or lowest levels of the construct of interest.

How common is construct truncation due to person sampling?

Construct truncation due to inadequate person sampling is prevalent in fields that rely on human participants such as psychology, epidemiology, and sociology. It occurs in spite of the best intentions of researchers because human participants are agents who cannot be

passively and randomly sampled, but, rather, take active parts in self-selecting into and out of research studies. This self-selection becomes problematic when related to the same constructs that are of interest in a study. It is has been observed, for example, that it is those individuals of the poorest health (Volken, 2013), lowest cognitive ability (Nishiwaki, Clark, Morton, & Leon, 2005), least conscientious and most neurotic personalities (Lönnqvist et al., 2007), or lowest and highest incomes (Bobko, Roth, & Bobko, 2001) who are most likely to decline to participate in research studies involving these respective constructs. In these instances, construct truncation can occur to varying degrees, depending on the extent to which individuals with the highest and/or lowest levels of that construct are under-represented in the sample. Evidence for the commonality of construct truncation comes from comparisons of research samples against population norms (Etter & Perneger, 1997), of respondents against non-respondents where information on the latter is available (e.g. Hill, Roberts, Ewings, & Gunnell, 1997), and of first assessments of participants who return to complete a follow-up study against those who drop out (Dollinger & Leong, 1993; Mein et al., 2012). To illustrate how widespread construct truncation due to person sampling is, possible examples across diverse research domains are outlined in Table 1.1. These are just a few examples, with others easily found in the literature. Unfortunately, it is difficult to characterise the overall scale of the problem because most selection is subtle and driven by forces not explicitly measured (Hunter, Schmidt, & Le, 2006). Furthermore, the population distribution of a construct is seldom known, making it difficult to ascertain whether and to what extent truncation has occurred in a given instance (Vink et al., 2004).

Table 1.1: Some examples of construct truncation due to person sampling

Construct	Nature of construct truncation	Possible implications
Family environment	Stoolmiller (1998) claimed that, due to selection procedures by adoption agencies, only the highest-quality family environments are represented in adoption studies.	The effects of family environments on various outcomes may be under-estimated in adoption studies because only the higher-quality family environments are represented
Disease risk	Miller (1994) noted that many studies of the association between type A behaviour and heart disease used only participants showing high levels of type A behaviour.	The relation between type A behaviour and heart disease may be under-estimated.
Intellectual disability and adaptive functioning	Murray, McKenzie, and Murray (2014) noted that studies attempting to estimate the association between adaptive functioning and intellectual ability have tended to use individuals with clinical diagnoses of intellectual disability.	The correlation between adaptive skills and intellectual ability may be under-estimated due to restricted range of intellectual ability because this diagnosis is made only when a client has an IQ<70.
Job selection tests	Schmidt, Shaffer & Oh (2008), LeBreton, Burgess, Kaiser, & Atchley (2003) and Sackett, Laczo, and Arvey (2002) discussed how job performance data on successful applicants is sometimes used to estimate the predictive validity of the selection tests and inter-rater reliability of job performance.	The predictive validity of selection tests and the inter-rater reliability of job performance are likely to be under-estimated when based on successful candidates alone.
Anti-social behaviour	Taylor (2004) found that twin families who showed higher levels of parental and child antisocial behaviour were less likely to respond to a mail survey.	Estimates of additive genetic variance and unshared environmental variance were inflated while estimates of shared environmental variance were attenuated.
Cognitive ageing	Deary, Gow, Pattie, & Starr (2012) demonstrated that two cohorts used to study cognitive ageing had higher means but smaller variances in IQ than the population from which they were sampled. Baseline cognitive ability also proved to be an important predictor of drop-out.	The effects of cognitive ability on later life outcomes and of various predictors on cognitive decline may be under-estimated.
Alcohol consumption	Heath, Madden & Marten (1998) found that participants in an alcohol challenge study had higher average alcohol consumption levels than the population from which they were drawn.	Individuals who would feel very intoxicated following an alcohol challenge were under-represented meaning that the genetic effects on subjective intoxication were likely slightly underestimated.

Construct truncation can also occur due to inadequate item sampling. A key goal of item selection is to include items that represent the full breadth of both *content areas* and *levels* of a trait (e.g. Diamantopoulos & Winkhofler, 2001; Murray, Eisner & Ribeaud, 2015). Both are important elements in the representativeness of items and in turn ecological validity; however, it is inadequate sampling with respect to the latter which creates construct truncation and on which the current thesis will, therefore, focus. Construct truncation due to inadequate item sampling is particularly prevalent in two kinds of study. The first is in studies concerned with constructs originating in psychopathological paradigms which have come to be studied in non-clinical populations. The second is when 'normal' measures are administered to extreme-scoring populations.

A range of traits related to the psychopathological constructs of autism spectrum disorder, psychosis, depression, and personality disorders are now routinely measured in non-clinical samples on the assumption that there is meaningful variation in these traits below their established clinical thresholds (e.g. Baron-Cohen, Wheelright, Skinner, Martin, & Clubley, 2001; Crawford & Henry, 2005; Jones & Paulhus, 2014; Verdoux & van Os, 2002). However, the scales measuring these constructs may not include items that can reliably capture this sub-clinical variation if they were developed and evaluated in the context of a psychopathological paradigm, and thus selected to show good discrimination near a clinical cut-off point and above (Reise & Waller, 2009). Items that showed good discrimination at moderate to low levels of that trait were, as a consequence, less likely to be selected. It is also common for scales to fail to include items that reliably capture variation at the clinical levels of a trait (Facon, Magis, & Belmont, 2011). For example, the gold standard intellectual ability assessment used to diagnose intellectual disability in children (the Wechsler Intelligence Scales for children, Fourth Edition; WISC-IV; Wechsler, 2003) shows marked floor effects (Whitaker & Gordon, 2012). There is evidence that at least one subtest is

frequently omitted because the child is unable even to understand the test instructions (Murray, McKenzie, & Murray, 2015).

Another area in which consideration of possible construct truncation is becoming increasingly pertinent is in the drive to produce briefer measures of constructs for contexts in which administration time is limited. For example, briefer versions of larger instruments are often of interest (e.g. Allison, Auyeung & Baron-Cohen, 2012; Donders, Elzinga, Kuipers, Helder & Crawford, 2013) for screening or reducing burden in clinical contexts. Similarly, brief measures of individual difference traits are appealing in large cohort or epidemiological studies (e.g. Rammstedt & John, 2007; see Weiss & Costa, 2014 for a criticism of this trend). However, abbreviating inventories will tend to entail a degree of construct truncation, especially if it further compounds selection on the basis of high item inter-correlations.

Evidence for construct truncation due to item selection comes from studies using an item response theory approach to examine the range of trait values for which a scale provides a reliable measure (Embretson & Reise, 2000). Item and test information is estimated for a range of trait values and portions of the continuum in which it - and by extension- reliability is low can be identified. Evidence also comes from score distributions where a disproportionately high number of scores are found to be at one or other end of the scale. Possible examples of construct truncation due to item sampling are provided in Table 1.2.

Table 1.2: Examples of construct truncation due to item selection

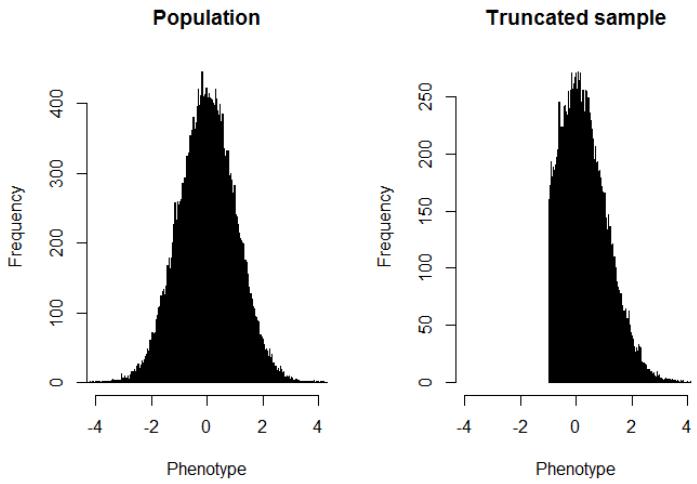
Construct	Nature of construct truncation	Possible implications
Personality	Gomez, Cooper, and Gomez (2005) found precision of measurement for the Behavioral Inhibition System/Behavioral Activation Systsm (BIS/BAS) scales (Carver & White, 1994) to be good only for moderately low to moderately high trait levels.	The scales cannot reliably capture high and low levels of the BIS/BAS traits.
Intellectual ability	Whitaker & Gordon (2012) demonstrated, using score distributions in low functioning individuals, that the Wechsler Intelligence Scale for Children Fourth Edition (WISC-IV) had marked floor effects.	Estimates of severity of intellectual impairment are poor for those with low IQ (also see Murray & McKenzie, 2014).
Physical functioning	Hays, Liu, Spritzer, & Cella (2007) showed that a physical functioning scale developed as part of a patient-reported outcome measurement and information system (PROMIS) project showed adequate reliability only at low levels of functioning.	Ability to detect the effects of interventions to improve physical functioning may be diminished by ceiling effects.
Memory	Uttl (2005) identified ceiling effects in a range of widely used memory tests when administered to normal healthy adults by examining test norms and in examining score distributions in a new sample.	Uttl (2005) listed the main issues as non-normal test scores, artificially lowered mean and attenuated standard deviations and validity.
Religiosity	Genia (2001) found evidence for ceiling effects in the Spiritual Well-being Scale (SWBS; Paloutzian & Ellison, 1982), especially in Christian respondents.	It is not possible to discriminate reliably between individuals scoring high on religious traits using the SWBS.
Criminal and deviant behaviour	Osgood, McMorris, and Potenza (2001) used an IRT approach to evaluate a self-report scale of offending behaviour from the Monitoring the Future study. The trait was measured reliably for moderate to serious offenders but not for the least delinquent third of the population	The scale cannot differentiate among individuals with low levels of criminal or deviant behaviour.
Emotional and behaviour problems	Van den Oord, Pickles, and Waldman (2003) used an IRT approach to show that a range of emotional and behaviour problem measures in the National Longitudinal Study of Adolescent were not measured well at the healthy end of the continuum	While the underlying liability for a psychopathological trait may be normally distributed, poor item sampling can lead to it appearing non-normal.

There is relatively less evidence bearing on the commonality of construct truncation due to item sampling. One reason is that until recently, most scales were developed and evaluated within a classical test theory approach which lends itself less well to identifying construct truncation. Classical test theory approaches typically assume that the reliability of a test is constant across the continuum of the measured trait and is, therefore, not equipped to identify reduced reliability at the extremes. Modern test theory, in dispensing with this assumption from the beginning, provides a more natural framework for evaluating possible construct truncation. Techniques from modern test theory are showing increasing uptake; therefore, more evidence on the matter is likely to emerge in the coming years. However, in tests developed using classical test theory it is still possible to apply IRT to the resulting item set and to examine score distributions for evidence of floor and ceiling effects.

What are the consequences of construct truncation for research?

The consequences of construct truncation can be considered in terms of how the sample distribution of a variable becomes distorted, relative to the corresponding population distribution. To illustrate, consider the distribution in Figure 1.1. On the left hand side is the population distribution for some phenotype (mean=0, SD=1). On the right hand side is the corresponding distribution in a sample in which truncation has occurred, specifically, trait levels below -1SD are not sampled.

Figure 1.1: Population and truncated sample distributions for a phenotype



Samples in which construct truncation has occurred will tend to mis-represent the mean of the target population. For example, in the phenotype represented in Figure 1.1, the mean in the truncated sample is not 0 but 0.29. Given that estimating the mean of a construct is rarely an interesting research goal in its own right, a more pressing problem is the mis-representation of mean *differences* between two groups when one group is subject to a greater degree of construct truncation than the other. Hunt and Madhyastha (2008) argued that many observed sex differences in intellectual ability were biased in this way. They noted that women are more likely than men to take the United States-based Scholastic Assessment Test for university admission. Given that taking the SAT is correlated with intellectual ability, they reasoned that the intellectual ability threshold for taking the SAT may tend to be lower for women. This would result in a higher proportion of less intelligent females taking the test, with the result that spurious group differences between males and females are introduced into many samples. Subsequent intellectual ability-related studies have noted how differential attrition across two groups being compared in longitudinal studies can also result in biasing of group comparisons (Dykiert, Gale, & Deary, 2009; Madhyastha, Hunt, Deary, Gale, & Dykiert, 2009).

Other examples come from possible diagnostic biases in psychiatry and clinical psychology whereby one group with a disorder is less likely to receive an appropriate diagnosis than another given the same level of impairment. When - as frequently occurs - a clinical diagnosis is in the inclusion criteria for an empirical study and those groups are then compared, this diagnostic bias can translate into inaccurate inferences with inappropriate theoretical implications (Krieser & White, 2014). For example, it has been argued that because of the preponderance of males with autism spectrum disorder (ASD) and the general perception that it is a 'male disorder', females must show more severe autistic symptoms or additional behavioural or psychological anomalies than males with equal levels of impairment to receive a clinical diagnosis (Murray et al., submitted). As a result, it is not clear whether the apparently greater severity of ASD symptoms in females compared with males with clinical diagnosis (Dworzynski, Ronald, Bolton, & Happé, 2012; Carter et al., 2007; Hartley & Sikora, 2009) is genuine or an artefact of a higher selection threshold - and greater construct truncation - at the point of diagnosis

This is just one specific example of the more general difficulty of testing the 'gender paradox' theory in clinical samples. The theory holds that whenever there are sex differences in symptom prevalence, the less-often affected sex will be the more severely affected sex. However, if the more-often affected sex is more likely to be selected into clinical samples given the same level of severity because of identification, referral and diagnosis biases, these samples will provide biased estimates of sex differences in symptomology (e.g. Biederman et al., 2014).

A related issue is 'Berkson's bias': the idea that estimates of psychiatric co-morbidity are inflated within clinical samples because different disorders may independently influence treatment-seeking (e.g. Maric, Myin-Germeys, Delespual, de Graaf, Vollenbergh & Van Os, 2004). Berkson's bias confounds attempts to estimate psychiatric co-morbidity from clinical

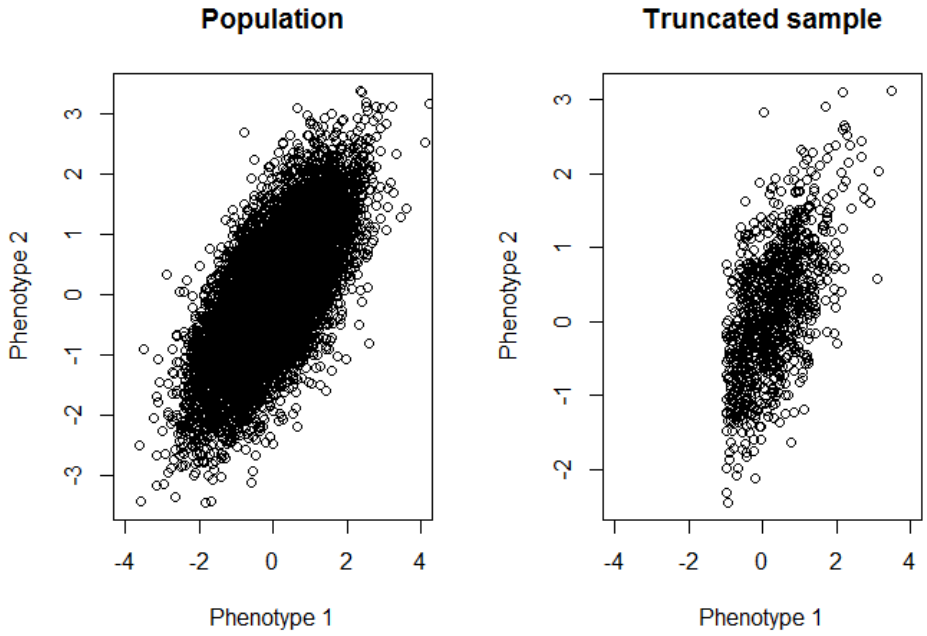
samples because those with the highest levels of general psychopathology (often meaning multi-morbidity) are over-represented.

Construct truncation also results in a reduction in the variance of the sample distribution of a construct relative to the corresponding population distribution, an effect often referred to as 'range restriction' (e.g. Alexander, Carson, Alliger, & Barret, 1984). For example, where the population variance for the phenotype shown in Figure 1.1 is 1 (left panel), the variance in the truncated sample is .63. The importance of this is less in the variance reduction per se than the knock-on effect it has on the covariances and correlations of that construct with others, both of which will be attenuated (Sackett & Yang, 2000; Ghiselli et al., 1981).

To illustrate, assume that the phenotype in Figure 1.1 is correlated with a second variable at r=.70 and that there is no additional truncation on this second phenotype. The scatterplot for these variables in the population is shown in the left panel of Figure 1.2 and in the truncated sample in the right panel of Figure 1.2. The correlation is attenuated in the truncated sample to r=.53 Examples of correlations possibly affected by range restriction abound in the psychological literature. Some of the best-studied examples are from organisational psychology in which selection tests are validated against job performance in successful applicants (e.g. Yang, Sackett & Nho, 2004). The correlations between selection test scores and job performance are attenuated in these samples because they exclude the individuals with lower scores. Furthermore, because variances, covariances and correlations represent the basic building blocks for more complex statistical tests, construct truncation biases these in much the same way. For example, in behaviour genetic models, range restriction may contribute to overestimation of heritability because MZ correlations will tend to be attenuated to a lesser degree than DZ correlations due to selective non-participation (Taylor, 2004); in factor analysis, it will tend to reduce factor loadings and inter-correlations

(Muthen, 1990); and in moderated multiple regression it will decrease the power to detect interactions (Aguinis, 1995).

Figure 1.2: Scatterplot for variables in population and truncated sample



More comprehensive discussions of the impact of range restriction on various statistics are available from several sources, including discussions of reliability (Fife, Mendoza, & Terry, 2012; Sackett et al., 2002), effect size measures (Bobko et al., 2001), regression (Cohen, Cohen, West, & Aitken, 2013), moderated multiple regression (Aguinis, 1995; Aguinis & Stone-Romero, 1997) factor models (Muthén, 1989, 1990), behaviour genetic models (Martin & Wilson, 1982; Neale, Eaves, Kendler, & Hewitt, 1989; Taylor, 2004; Dominicus, Palmgren, & Pedersen, 2006), and meta-analysis (Hunter et al., 2006).

Finally, construct truncation can result in sample distributional shapes that depart from that of the underlying population distribution. Constructs that are normally distributed

in the population will often be skewed in a sample subject to construct truncation. Although change in shape can result in a slight bias in the correlation between that construct and others (e.g. Bishara & Hittner, 2012), the bigger problem is that it can lead to masked or spurious interactions when the construct is used as an outcome in, for example, moderated multiple regression models, gene-environment interaction models, or factorial ANOVA or other models employed to test higher-order effects (Embretson, 1996; Kang & Waller, 2005; Molenaar & Dolan, 2014; Schwabe & van den Berg, 2014; Stone & Holllenbeck, 1989). For example, Wang Zhange, McArdle, & Salthouse, (2008) illustrated that when a test exhibits ceiling effects but groups differ in proportions of participants scoring at ceiling, spurious group by time interactions can be observed. They compared memory test scores from older and younger adults measured repeatedly over time. As the latter had more participants scoring at ceiling at baseline, they exhibited the smallest change; however, because their initial were scores likely under-estimated due to ceiling effects, the interaction between age and time was likely to be at least partly spurious.

However, the fact that a construct is truncated does not, by definition, mean that there is serious bias in statistical analyses involving it. As discussed in later chapters, the degree and manner to which this occurs depend on the statistical test of interest, the population to which one wishes to generalise and, of course, the extent of construct truncation.

What can we do about construct truncation?

Given the bias in statistical results that can arise due to construct truncation, it is important to consider what can be done to identify, characterise and mitigate such bias. Identifying problematic construct truncation requires some knowledge of the population distribution of the construct. Occasionally this is possible because a measure has been normed using population data. However, as Marcus & Schütz (2005) caution, even norming

data can be subject to degrees of truncation, especially when it relies on volunteer samples. Other times, a reasonable assumption can be made about the population distribution. In the cases where the population distribution is known, or can be assumed, it may be possible to diagnose and even correct for bias due to construct truncation. For example, information and assumptions about the population distribution of a trait or its relation to selection into or out of the sample can be used to create selection models, derive sampling weights, apply range restriction formulae, or employ models such as tobit regression models to attempt to correct for the construct truncation. These corrections are, however, fallible, require strong assumptions, difficult to apply retrospectively to already-published results and become impractical for more complex selection scenarios and statistical models.

For example, a variety of different range restriction corrections tailored to different scenarios are available to estimate the correlation between variables in the population given its value in a selected sample (see Hunter et al., 2006; Sackett & Yang, 2000; Schmidt, Oh & Le, 2006) . Reversing the corrections also allows an estimate of the opposite, which may be useful, for example, in studying the effects of range restriction in methodological studies. To obtain accurate results using a range restriction correction is, however, very difficult. It is, first of all necessary to have knowledge of the kind of selection that has occurred. A distinction can be made, for example, between 'direct' and 'indirect' selection, with the former referring to selection based on the observed scores of one of the variables to be analysed and the latter referring to selection on some third variable that is correlated with these scores. Applying a direct range restriction correction to indirectly selected variables will generally lead to under-correction (Alexander et al., 1984; Linn, 1983). The selection mechanism underpinning a given case of range restriction is not something that can be inferred from the sample data alone (e.g. see Hanges, Rentsch, Yusko & Alexander, 1991)

and in the absence of information about the true selection mechanism, there has been a tendency in empirical studies to assume simple direct selection (Sackett et al., 2007).

Second, it is often difficult to obtain good estimates of the values required to apply range restriction corrections. For example, the correction formula for the simple case of direct range restriction (known as Thorndike case II) is:

$$r_{XY} = \frac{r'_{XY}\left(\frac{SD_X}{SD'_X}\right)}{\sqrt{1 - r'^{2}_{XY} + r'^{2}_{XY}\left(\frac{SD_X^{2}}{SD'^{2}_X}\right)}} \tag{1.1}$$

where r_{XY} is an estimate of the population correlation, r'_{XY} the correlation in the selected sample, is the ratio $\frac{SD_X}{SD'_X}$ and $\frac{SD_X^{2}}{SD'^{2}_X}$ are standard deviation and variance ratios respectively for X in the population to sample and thus requires estimates of the population variance of the variable to be known. As noted above, however, the population variance of many variables is unknown and even when normative data is available this may itself have been subject to a degree of selection and, therefore, not provide an accurate estimate of population variance (Marcus & Schuz, 2005).

Finally, this and other range restriction correction formulae assume that the regression of Y on X (or of each variable on all others in multivariate extensions) is linear and homoscedastic across its entire range. This assumption is not testable for the parts of the distribution that are unmeasured. In fact, given the frequency with which hypotheses regarding non-linear effects of the kinds of individual difference traits that tend to influence selection into research studies are advanced, it is likely that at least mild violations of this assumption occur from time to time. Miller (1994), for example, noted that treatment effects in high risk studies in behavioural medicine (which are subject to range restriction on the

outcome of interest) are likely to be non-linear because treatment will be most effective for the most severely affected.

Therefore, there remains an important role for simulation studies modelling the mechanisms and consequences of construct truncation to characterise its effects and develop and test new ways to counteract it. Finally, outside of a few specific research areas (particularly personnel selection), construct truncation has received little attention and it seems that there is a general lack of awareness of its effects. Therefore, it is important to continually assess, both conceptually and statistically, whether, and in what way, the empirical evidence in a given research domain is affected by construct truncation. It is based on these observations that the aims of the current thesis were developed: to contribute to the characterisation and mitigation of the adverse effects of construct truncation on psychological research.

I begin in **Chapter 2** with a focus on construct truncation due to person selection. I outline the potential issues of relying on clinically ascertained samples when conducting research on psychopathological phenotypes. I use a specific example from an empirical research domain concerned with understanding the etiology of autism spectrum disorders (ASD). I use this example to address how relying on clinical samples for empirical research into particular disorders may provide distorted pictures of the inter-relations of the symptoms of those disorders, as well as their relations with putative causes and consequences. I argue that when a clinical trait is merely the extreme end of a continuum, clinical samples can be subject to strong and distorting selection. I argue that the low inter-correlations among the symptoms of ASD cited as evidence that they have distinct etiologies may be under-stated because this evidence has been based on clinical samples which are by definition selected on these symptoms. I support these arguments with evidence from a statistical simulation and a real data example.

In **Chapter 3** I argue that construct truncation due to person selection can work in subtle, hard-to-detect ways, drawing on an example from individual differences research. I discuss the 'intelligence compensation hypothesis'. This is the idea that conscientiousness levels become calibrated to intellectual ability levels because less intelligent individuals need to work harder to achieve in life, while more intelligent people can rely on their superior abilities. I argue that much of the evidence for the 'intelligence compensation hypothesis' could reflect artifacts of using samples selected with respect to achievement. Specifically, I argue that many of the negative associations between conscientiousness and cognitive ability that have been observed in these samples have likely been spurious because of the actions of a 'composite selection' mechanism. Composite selection in this context refers to self-selection of participants higher in achievement into the populations subsequently used as samples in research studies. Here, a 'composite' of conscientiousness and cognitive ability determines achievement (which influences later selection into a research study). Because selection occurs on a composite of conscientiousness or cognitive ability rather than either alone, high levels of cognitive ability can compensate for low levels of conscientiousness, and vice versa with regards to entry to the research sample. Samples of individuals selected on achievement can, thus, show a negative correlation between these two variables, even if the two constructs are completely independent at the population level. I evaluate this hypothesis using a real data example. In the example, I artificially introduce selection on achievement in the sample to investigate how this affects the apparent association between conscientiousness and ability. Results suggest that the true association between cognitive ability and conscientiousness may be zero or positive in reality but that construct truncation on achievement can give the appearance of a negative association.

In **Chapter 4,** I focus on construct truncation due to item selection. I discuss an important consequence of selecting items that fail to measure the full range of a construct

reliably: that of potentially distorted estimates of GxE interaction in behaviour genetic models. I outline the implications of using the raw scores from scales which have poor discrimination at one end of their continua. I discuss two potential solutions: a non-linear transformation of raw scores and an item response theory score from an appropriate measurement model. I then use statistical simulation and examples in real data to explore the extent to which these proposed solutions mitigate the adverse effects of using a scale with poor discrimination in one part of the latent continuum. Results suggest that transformed raw scores and IRT scores perform reasonably well in reducing the bias that would otherwise be introduced into tests of GxE when using scales that fail to measure one end of a trait distribution reliably; however, neither method eliminated bias. Of the two methods, I recommend using IRT scores because they showed slightly less bias and type 1 and type 2 error rates were slightly better than those using a non-linear transformation. I also discuss other advantages of using IRT scores over transformed raw scores.

Finally, in the **Chapter 5, the Discussion,** I show how the issues raised in the 3 previous chapters are all aspects of the broader issue of construct truncation. I discuss the collective implications of the thesis studies for the mechanisms of, consequences of, and possible solutions to construct truncation. Specifically, I argue that the results of the studies in the preceding chapters suggest that distortions in commonly used analyses due to construct truncation in varying forms and degrees of severity are likely to occur in a broad range of empirical research studies. I discuss the circumstances under which it tends to occur and the kinds of misleading substantive conclusions that can result. I argue that it is important to consider the possibility that data are affected in this way and to take steps to characterise and mitigate the consequences. I also highlight the limitations of the research contained within the current thesis and offer some suggestions for future research. I end by summarising the most

promising solutions suggested by the research contained within the thesis but also highlight the challenges of implementing them in practice.

Chapter 2: Construct truncation due to the use of clinically ascertained samples in autism spectrum disorder research

In research into clinical phenotypes, it is common to recruit samples comprised solely of individuals who have received a formal diagnosis of that disorder. It is similarly common to analyse clinically diagnosed and control cases separately. However, the assumption that clinical phenotypes are on continua that span clinical and 'normal' levels is gaining increasing acceptance (e.g. Caspi et al., 2014). In this view, individuals who receive a clinical diagnosis do not differ qualitatively from those without; they merely represent an extreme on the same continuum. This implies that to obtain unbiased estimates of population parameters concerning that phenotype, clinical cases and controls should be analysed together, irrespective of diagnostic status. Reliance on clinically-diagnosed samples would restrict the phenotypic ranges studied and, in turn, attenuating correlations involving these phenotypes. In this chapter, I discuss how this kind of range restriction may have affected our understanding of the inter-relatedness of the classical triad of autism spectrum disorders (ASD).

The use of clinical and control samples in autism spectrum disorders

There is increasing consideration and acceptance of the hypothesis that ASD traits exist on continua that span clinical and non-clinical levels (e.g. Frazier et al., 2010; Lundström et al., 2012; Murray, Booth, McKenzie, Kuenssberg, & O'Donnell, 2014). In this view, individuals who receive a clinical diagnosis of ASD are simply the extreme end of this continuum, rather than manifesting some qualitatively distinct condition (Austin, 2005). This being true, any correlation-based analysis that focuses exclusively on either clinically diagnosed or control individuals is liable to yield phenotypic associations that are attenuated due to range restriction. However, to analyse clinically diagnosed and control individuals as a

single sample is rare. Most studies focus on one group, or - where both are recruited - conduct separate analyses by group (e.g. Baron-Cohen & Wheelright, 2004; Wheelright et al., 2006; Stewart, Allison, Baron-Cohen, & Watson, 2015). In the section that follows, I discuss possible implications that this has had for the hypothesis that ASD is a 'fractionable disorder'.

The fractionable triad hypothesis

There is considerable interest in establishing how strongly related different areas of deficits in ASD are. The question has formed the basis of numerous empirical and review studies, including a recent special issue of the journal *Autism* (2014, vol 18, issue 1). While it has long been acknowledged that ASD is an extremely heterogeneous disorder (Rutter, 2014), over time these observations have evolved into the hypothesis that ASD is fractionable disorder; that is, it comprises multiple, somewhat independent, symptom domains (see Happé & Ronald, 2008 for a review).

When expressed in terms of the classical triad of ASD the hypothesis is referred to as the 'fractionable triad' hypothesis. The hypothesis holds that the three classical symptoms of ASD: deficits in reciprocal social interaction, communication, and restrictive and repetitive stereotyped behaviour, are not all manifestations of the same underlying disorder; rather they represent separate domains of impairment whose confluence is ASD. From this basic idea, discussions have expanded to consider the fractionation of ASD symptomology more broadly; in terms of the two diagnostic domains of the DSM 5 (Mandy et al., 2014); cognitive symptoms (Brunsdon & Happé 2014); and genetic and environmental etiology (Dworzynski et al., 2009; Mazefsky et al., 2008; Robinson et al., 2012).

The fractionation hypothesis has received so much attention because it is viewed as having important theoretical and practical implications**.** First, it underscores the importance

of using assessments that capture all symptom domains because if these domains are relatively independent, global assessments may omit key features of an individual's symptom profile (Happé & Ronald, 2008). Second, it implies no requirement for ASD symptoms to be specific to ASD because, under the fractionation hypothesis, ASD is the co-occurrence and not the root cause of specific ASD symptoms. Third, distinct etiologies of ASD symptoms suggest that searches for specific causes should focus efforts on specific symptoms. A fourth possibility is that treatments will have symptom-specific rather than global effects and, by the same token, should be targeted at specific symptoms to maximise chances of alleviation.

Historically, key pieces of evidence contributing to development of the fractionation hypothesis and now cited in its support are correlations between different symptom domains in individuals with a clinical diagnosis of ASD that are only low to moderate (Brunsdon & Happé, 2014; Dworzynski et al., 2009; Happé & Ronald, 2008 ; Kolevzon et al., 2004; Mandy et al., 2014). However, symptom correlations in clinically diagnosed individuals may represent significant underestimates of the corresponding population values because of range-restricting selection arising from the diagnostic process. In the section that follows, I use what is known about the diagnostic process and subsequent use of clinically diagnosed samples in research to build a statistical model of range restriction in fractionation hypothesis research. I use this model to gauge the impact of this practice, and provide a range of estimates for the 'true' associations between symptoms.

Individuals who meet the diagnostic criteria for ASD are a select group comprising approximately only 1% of the population (Baird et al., 2006; Baron-Cohen et al., 2009). They are not a random sample, but a select sub-section of the population representing the extremes of ASD traits. It has long been known that when samples are selected with respect to some trait, the variance of that trait is attenuated (e.g. Pearson, 1903). This is range restriction, and it tends to reduce correlation with other variables as well. The simplest form of range

restriction is 'explicit' or 'direct' selection on some variable X, when the correlation between X and some other variable Y is of interest. That is, the variable X on which the sample is selected is identical with the variable X which is utilised in analyses in the selected sample, and there is a strictly observed cut-off score for inclusion. This situation most commonly arises when X is some aptitude test used to select candidates for a job and Y is an index of job performance in the successful candidates in order to validate the aptitude test X (e.g. Berry et al., 2011). The extent to which the variance of X and its correlation with Y is reduced depends on the strength of selection, which can be quantified using the ratio $\frac{SD\prime}{SD}$ = u_X, where SD' is the standard deviation of X in the selected sample and SD is the standard deviation of X in the population. Lower values of u_X represent stronger selection on the variable X. Given u_X, the Pearson correlation between X and Y in the selected group will be:

$$r'_{XY} = \frac{u_X r_{XY}}{\sqrt{u_X{}^2 r^2{}_{XY} + 1 - r^2{}_{XY}}} \quad , \tag{2.1}$$

where r_{XY} is the correlation between X and Y in the unrestricted population. From Eq. 2.1, it can be seen that whenever u_X is less than 1, the Pearson correlation between X and Y will be downwardly biased in the selected sample.

Diagnosing ASD is also a selection process that reduces variance in the traits of interest and, in turn, is likely to attenuate symptoms correlations relative to the population. The selection process is more complex than the job selection example and cannot be represented using the simple model in Eq. 2.1. First, ASD diagnosis does not involve direct selection on measured scores on the X variable(s). That is, a clinician cannot simply 'read off' an individual's levels of, say, social, communication and restrictive, repetitive

impairments or behaviours and assign those with a score above certain cut-off points a diagnosis of ASD. Instead, the process involves a combination of formal assessment and clinical judgement (Allison et al., 2012). As a consequence, scores on measures of ASD symptoms obtained in subsequent research will not be identical to the criteria by which a clinician has assigned a diagnosis. The process of diagnosing ASD and then selecting participants for a research study, therefore, represents an example of 'indirect selection', defined as occurring when the selection variables are not identical with the variables that form the basis of subsequent empirical analyses (Hunter et al., 2006). In the terminology of range restriction, therefore, the triad or other features of ASD of interest in an empirical study represent 'incidental variables' which are selected by virtue of being correlated with the unmeasured variables on which selection (diagnosis) takes place (i.e. the selection variables).

Another way in which the case of ASD diagnosis is more complicated than the simple job selection example is that ASD diagnosis requires the presence of symptoms in *multiple* domains to be present. This makes ASD diagnosis a case of simultaneous multivariate – rather than univariate - selection (Sackett & Yang, 2000). DSM IV diagnosis required deficits in three of the areas of the classical triad and was, therefore, a case of trivariate selection. The new DSM 5 criteria requires deficits in social communication – which combines the social and communication dimensions of the classical triad - coupled with the presence of restricted, repetitive behaviours, entailing a shift to bivariate selection. A multivariate selection formula was developed by Aitken (1934) and subsequently extended by Lawley (1944) to deal with situations such as this in which samples are selected on multiple variables. The formula provides a correction to estimate population associations in range-restricted samples. Based on the formula, an estimate of the population variance-covariance matrix **V** of the selection and incidental variables in the population can be obtained using:

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V_{p,p}} & \boldsymbol{V_{p,p}V'_{p,p}V'_{p,n-p}}^{-1} \\ \boldsymbol{V'_{n-p,p}V'_{p,p}}^{-1}\boldsymbol{V_{p,p}} & \boldsymbol{V'_{n-p,n-p}} - \boldsymbol{V'_{n-p,p}}(\boldsymbol{V'}^{-1} - \boldsymbol{V'_{p,p}}^{-1}\boldsymbol{V_{p,p}V'_{p,p}}^{-1})\boldsymbol{V'_{p,n-p}}^{-1} \end{bmatrix} \quad (2.2)$$

where $\boldsymbol{V_{p,p}}$ is the variance-covariance matrix of the *p* selection variables in the population, $\boldsymbol{V'_{p,p}}$ is the variance-covariance matrix of the *p* selection variables in the selected sample, $\boldsymbol{V'_{n-p,p}}$ and $\boldsymbol{V'_{p,n-p}}$ are the covariance matrices for the *p* selection variables and *n-p* incidental variables in the selected sample and $\boldsymbol{V'_{n-p,n-p}}$ is the covariance matrix of the *n-p* incidental variables in the selected sample. However, it is apparent from Eq. 2.2 that in order to apply this correction, it is necessary to have information on the selection variables. This is simply not available in the case of ASD diagnosis because, as mentioned above, the selection variables are a composite of formal assessment and clinical judgement and the latter is not directly quantifiable. In fact, this information is rarely available for any multivariate selection problem (Hunter et al., 2006). This creates a challenge with respect to estimating the degree to which symptoms of ASD cluster together because any sample restricted to individuals with ASD will be liable to under-estimate their association, but owing to a lack of information on the selection variables, it will be difficult to assess the extent of the bias.

While the possibility that range restriction may undermine the validity of results from clinical ASD samples has been noted (Happé & Ronald, 2008), there has not as yet been any systematic study or attempt to quantify the consequences of this kind of selection. ASD is fundamentally a clinical disorder and inferences regarding ASD should, therefore, come at least in part from samples of individuals who are actually diagnosed with the disorder: it would be undesirable to disregard all studies restricted to individuals with diagnosed ASD

from consideration because they are affected by range restriction. It was, therefore, the aim of the present study to attempt to characterise and quantify the effects of simultaneous selection within research studies focussed on individuals with a clinical diagnosis of ASD. I first present the results of a brief simulation exploring the potential effect of simultaneous selection on estimates of the inter-correlation among symptoms of ASD. I then provide a real data example comparing the correlation among ASD symptoms in individuals with diagnoses of ASD to a combined sample which includes both individuals with and without diagnoses of ASD.

Method

Simulation study

To receive a diagnosis of ASD, an individual has traditionally had to show deficits in all three domains of the classical triad, which I will here abbreviate to 'Soc', 'Comm' and 'RSB'; therefore, the majority of samples of individuals with clinical diagnoses of ASD are selected on these three traits. This makes it reasonable to hypothesise that it is this selection that produces that relatively low correlations among them that has inspired the 'fractionable triad' hypothesis. I, therefore, explored the effects of simultaneous selection using a range of possible population correlation magnitudes among simulated Soc, Comm and RSB variables. All analyses were conducted in R statistical software (R Core Team, 2013).

Population model

I began with a model in which RSB, Comm and SS had a trivariate normal distribution with means of zero and variances of 1 in the population. This corresponds to the idea that the traits are normally distributed in the population, with ASD representing the extremes of these traits (e.g. Austin, 2005; Lundström et al., 2012). I simulated 10,000,000

cases to represent this population. Across different simulation conditions I varied the population correlations between RSB, Comm and SS. The population correlations utilised are provided in Table 2.1. Reflecting the evidence that Comm and SS are more strongly inter-related than either is with RSB, I simulated non-uniform population correlations among the triad (e.g. Dworzynski et al., 2009).

Selection model

I simulated simultaneous selection from the populations described above in a manner designed to mimic the diagnostic process. I did this by selecting cases from the uppermost part of the univariate distributions of the three variables. I did not select on RSB, Soc and, Comm scores directly but generated a 'selection variable' for each. These selection variables were correlated at $r \approx 0.75$ with the corresponding symptoms to represent indirect selection and an appropriate level of fallibility of the diagnostic process. I then selected cases based on being above the 95th percentiles on these selection variables. The 95th percentile has been used to define abnormality in studies of ASD traits utilising general population participants (e.g. Robinson et al., 2012).

Quantifying bias

I evaluated the effect of simultaneous selection on the sample symptom inter-correlations and quantified the degree to which these sample estimates under-estimate the corresponding population value using percentage bias, computed as:

$$(r' - r)/r \times 100\%$$

where r is the simulated population correlation and r' is the correlation in the selected sample. Percentage bias is commonly used to evaluate the extent to which a parameter is estimated accurately in simulation studies, including those concerned with the effects of

range restriction (e.g. Le & Schmidt, 2006). I also provide the determinant of the covariance matrix for the three variables before and after simultaneous selection. This provides a measure of generalised variance. The percentage bias in this statistic is computed in an analogous way to the bias in the individual correlation coefficients.

Real data example

To complement the simulation study, I provide a real data example in which I compared the correlations between ASD symptoms in clinically diagnosed individuals to the corresponding correlations in a combined sample of individuals with a diagnosis of ASD and controls.

Measures

I used the Autism Spectrum Quotient (AQ: Baron-Cohen et al., 2001). The AQ is a 50-item questionnaire (50% reverse-keyed) assessing 5 domains: Social Skill, Attention Switching, Attention to Detail, Communication and Imagination. Item content is based on the classical triad of ASD as well as other cognitive traits associated with ASD. Items are scored on a dichotomous response scale resulting in a possible range of scores for each domain from 1-10.

Previous studies have suggested that the AQ has favourable psychometric properties including good test-retest reliability, acceptable internal consistency, higher scores in clinically diagnosed than control samples, normally distributed scores in the population and correlations with other features of ASD (e.g. Allison et al., 2012; Baron-Cohen et al., 2001; Takagishi et al., 2010). The advantage of the AQ in the context of the current study is that is based on a dimensional approach to ASD which conceptualises ASD traits as continuous in the population and, therefore, measurable even in individuals who do not meet diagnostic

criteria for ASD. Moreover, it was specifically designed to measure ASD traits across a broad range from normal to clinical levels. Indeed, evidence suggests that the AQ successfully captures variation in ASD traits in both clinically diagnosed and non-ASD individuals (Baron-Cohen et al., 2001; Hoekstra et al., 2011; Wheelright et al., 2010).

Participants

Controls

Control participants came from 2 sources. Ninety eight participants (27 males, 70 females and 1 'other' gender) came from an ongoing study of emotion recognition and ASD traits which included the AQ as a measure. The 98 were selected from a broader pool of participants that also included individuals who self-reported a diagnosis of intellectual disability, ASD or another psychiatric disorder. The mean age of this sub-sample was 31.0 (SD = 12.5). The majority reported their occupation as 'student'. An additional 132 participants (27 males, 105 females) came from an ongoing study of sex differences in ASD traits. The mean age of this sub-sample was 27.7 (SD = 12.7). Sixty eight of these participants reported their occupation as 'student'. Both of the studies above had received ethical approval from the relevant ethics body. Participants were in both cases recruited online and from the University community. Online recruitment was via social networking sites such as facebook and twitter as well as dedicated study participation sites.

Cases

Participants with ASD came from a previous study of the AQ in clinically diagnosed individuals. The sample has been utilised and described in previous publications (Booth et al., 2013; Kuenssberg et al., 2014; Murray et al., 2014) and is described comprehensively in Kuenssberg et al. (2014). The full sample includes 148 participants (107 males and 41

females) with a diagnosis of Asperger Syndrome or high functioning autism. High functioning autism was defined as meeting the criteria for autism but having normal intellectual functioning. Asperger syndrome was defined as meeting the criteria for high functioning autism but with no history of language delay. The mean age of the sample was 33.3 (SD = 10.7). In the current study, I used the subset of participants with complete data on the five domains measured by the AQ (N = 132-135). As the data were fully anonymised prior to receipt it is not possible to identify the specific demographic composition of this sub-sample.

Statistical Procedure

I first computed Pearson correlations between the 5 AQ domain scores in the ASD group and in a sample that combined both the ASD and control participants. I quantified the difference in Pearson correlation between the whole sample and ASD sub-sample, in a similar way as in the simulation study by computing the percentage difference between whole and ASD sub-sample:

$$(r_{ASD} - r)/r \times 100\%,$$

where r_{ASD} is the correlation in the ASD sub-sample and r is the correlation in the whole sample. I also computed an estimate of internal consistency for each of the AQ domains in the whole sample and the case sub-sample using Cronbach's alpha.

Finally, I applied the most commonly used method of range restriction correction to the correlations in the case sample: the Thorndike case II formula from Eq. 1.1. I used this to obtain *r**, an estimate of the range restriction corrected correlations assuming that the combined sample was the population. Although this is technically not the correct formula because it is designed for a situation in which direct selection has taken place, in practice it is

often used because of insufficient information about the selection process to permit the application of Eq. 2.2. The difference between results obtained using Eq 1.1 and the empirical estimates of the correlations are evaluated using the percentage difference between the range-restriction corrected correlation and the whole sample correlation. For SD' I used the empirical estimate of SD in the case sample; for SD I used the empirical estimate of SD in the combined sample; and for *r*' I used the empirical estimate of *r* in the case sample. The purpose of this analysis was to evaluate the extent to which the most commonly used method of range restriction correction would give similar results to using the relevant population to which the formula aims to correct to.

Results

Simulation study

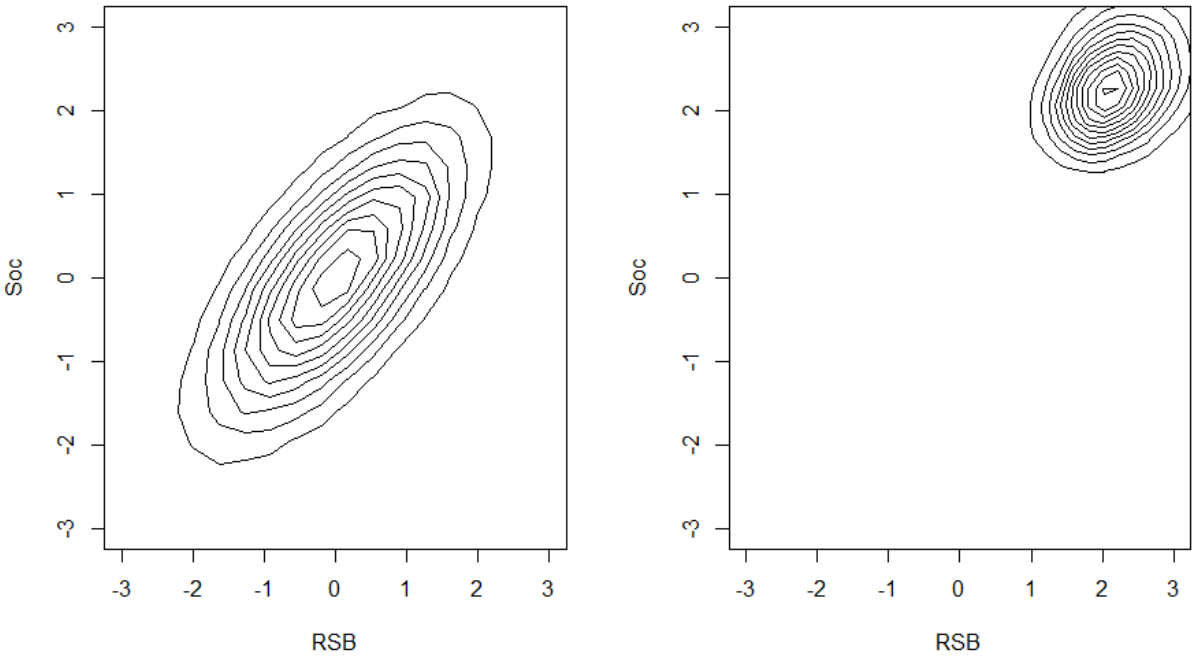
Results from simulating simultaneous selection on Soc, Comm and RSB are provided in Table 2.1. These show how a selection mechanism representing the ASD diagnosis can lead to substantial under-estimates of symptom inter-correlations in samples of clinically diagnosed individuals. For example, if the simulation selection mechanism proposed has successfully represented a situation close to reality, then an observed correlation between RSB and Soc of $r = .26$ could correspond to a population correlation of $r = .60$. Other possible magnitudes of population and corresponding sample correlations can also be read off from Table 2.1.

The results also highlight how the biggest percentage under-estimates of the correlation among symptoms occur when the relevant population correlation is itself smaller. For example, the percentage bias for a population correlation of .95 was only -13% whereas the percentage bias for a population correlation of .40 was approximately -65%. Therefore, to the extent that simultaneous selection attenuates symptom inter-correlations in ASD samples,

it is likely to do so to a greater extent for those domains that have smaller population correlations.

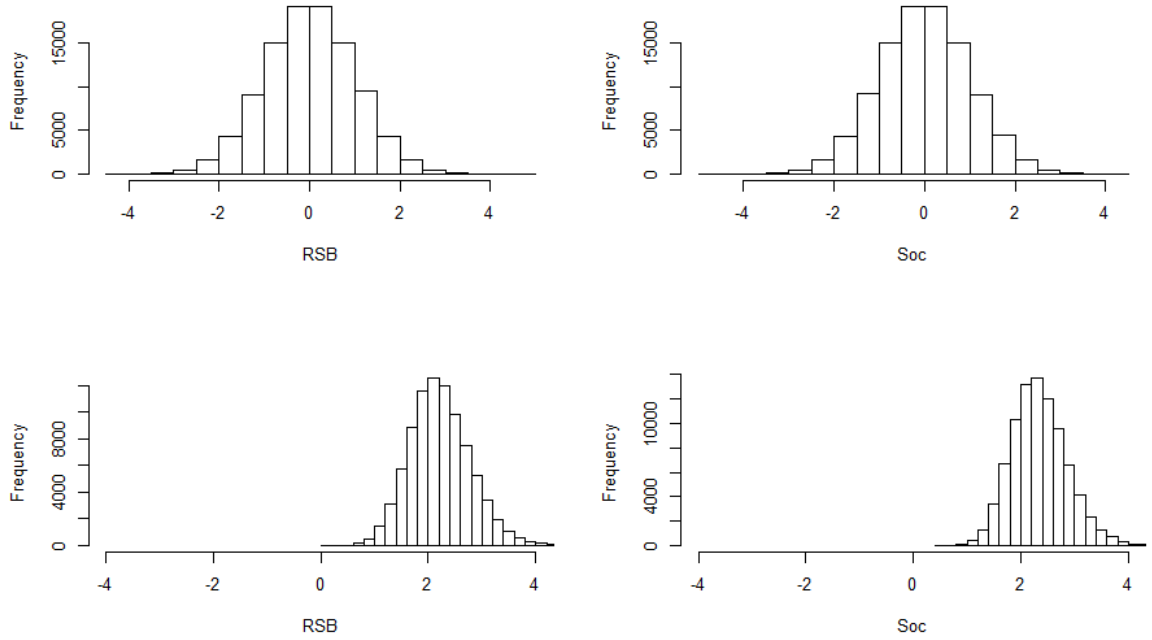
To provide more in-depth example, Figure 2.1 shows the bivariate density of RSB and Soc in the condition from the first row of Table 2.1. The contour plot in the left panel shows the population distribution and the right panel shows the distribution after simultaneous selection. It can be seen that the location of the distribution shifts towards the clinical end, the variance is reduced and the elliptical shape of the population distribution is lost in line with the decreasing correlation between the variables.

Figure 2.1: Contour plots of RSB and Soc before and after selection



.

Figure 2.2: Marginal distributions of RSB and Soc before and after selection



The marginal distributions of the same two variables are shown in Figure 2.2. These, in addition to showing the shift in location and reduction in variance, highlight the positive skewness introduced by simultaneous selection. The top row shows the distributions in the population and the bottom row shows the distributions after simultaneous selection. The skews of RSB and Soc after selection are .35 and .41 respectively.

Table 2.1: Extents of attenuation of symptom inter-correlations and generalised variance under simultaneous selection

RSB- Soc			RSB -Comm			Comm-Soc			Generalised variance			Prevalence
R	***r′***	**% bias**	***r***	***r′***	**% bias**	***R***	***r'***	**% bias**	**|∑|**	**|∑|'**	**% bias**	
.70	.34	-51	.70	.35	-50	.95	.83	-13	.049	.006	-87	.86%
.60	.26	-57	.60	.26	-57	.95	.83	-13	.062	.007	-88	.69%
.50	.19	-62	.50	.20	-60	.95	.83	-13	.073	.008	-89	.54%
.40	.13	-68	.40	.14	-65	.95	.83	-13	.082	.008	-90	.42%
.30	.10	-67	.30	.10	-67	.95	.83	-13	.089	.010	-91	.32%
.60	.26	-57	.60	.25	-58	.90	.70	-22	.118	.013	-89	.65%
.50	.19	-62	.50	.19	-62	.90	.70	-22	.140	.014	-90	.51%
.40	.13	-68	.40	.13	-68	.90	.71	-21	.158	.014	-91	.40%
.30	.09	-70	.30	.09	-70	.90	.71	-21	.172	.014	-92	.30%
.70	.35	-50	.70	.34	-51	.80	.49	-39	.164	.019	-88	.73%
.60	.25	-58	.60	.24	-60	.80	.50	-38	.216	.023	-90	.57%
.50	.18	-64	.50	.18	-64	.80	.51	-36	.260	.024	-91	.45%
.40	.14	-65	.40	.14	-65	.80	.52	-35	.296	.025	-92	.36%
.30	.10	-67	.30	.11	-63	.80	.51	-36	.324	.025	-92	.26%
.60	.26	-57	.60	.26	-57	.70	.37	-47	.294	.028	-91	.51%
.50	.19	-62	.50	.18	-64	.70	.37	-47	.360	.031	-92	.40%
.40	.14	-65	.40	.14	-65	.70	.39	-44	.414	.031	-93	.30%
.30	.10	-67	.30	.10	-67	.70	.39	-44	.456	.031	-93	.22%
.50	.19	-62	.50	.19	-62	.60	.29	-52	.440	.036	-92	.35%
.40	.14	-65	.40	.14	-65	.60	.29	-52	.512	.037	-93	.26%
.30	.09	-70	.30	.10	-67	.60	.21	-65	.568	.038	-93	.19%
.40	.14	-65	.40	.14	-65	.50	.21	-58	.590	.040	-93	.23%
.30	.08	-73	.30	.09	-70	.50	.20	-60	.660	.041	-94	.17%

Note. RSB= restricted, repetitive behaviours, Comm= communication, Soc= social skills.

Real data example

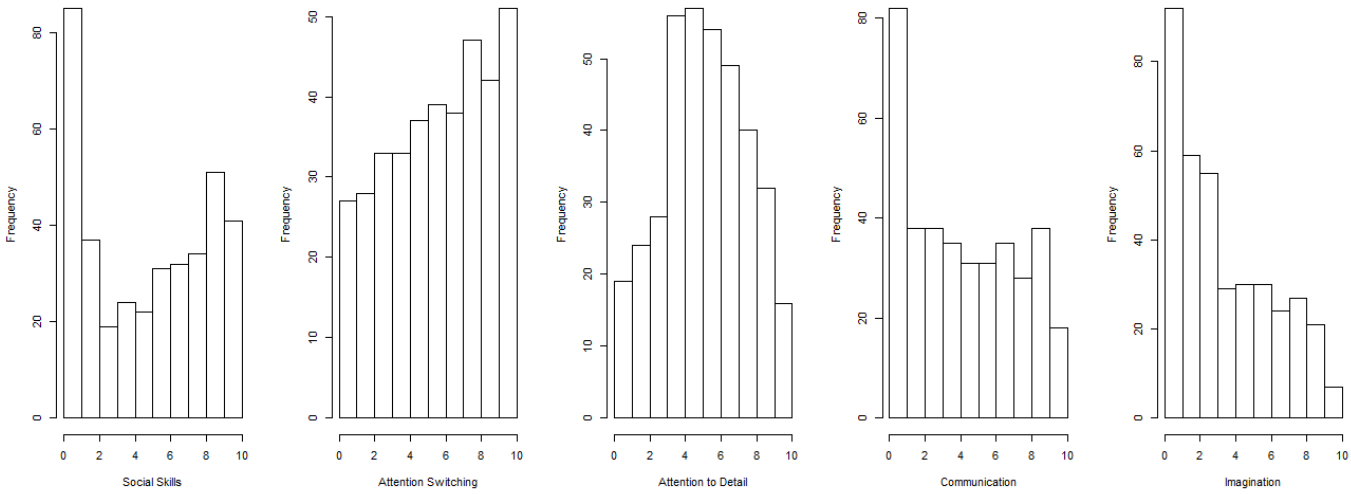
Descriptive statistics for the cases, controls and combined sample are provided in Table 2.2. As expected, the mean scores for all 5 domains were higher in the cases than in the controls. The standard deviations did not differ markedly between the cases and control groups but, as expected, were larger in the combined sample than in either of the case or control sub-samples. The ratios of the SDs of the domain scores in the ASD sample to those in the combined sample are in the last column of Table 2.2. The largest SD difference was observed in Attention Switching domain (.52). The smallest difference was for the Attention to Detail domain (.97) and suggested only minimal range restriction.

Figure 2.3 shows the score distributions in the combined sample. There is not much evidence for bi-modality arising from clinical and control samples. Bi-modality would suggest possible range enhancement. However, the Social Skills, Communication and Imagination subscales showed some evidence of floor effects while the Attention Switching subscale showed some evidence of ceiling effects. Figure 2.4 shows the scatterplot matrix of the AQ subscale scores in the combined sample. Loess lines are added to highlight any non-linearity of the associations. Such non-linearity could indicate that the assumptions of range restriction corrections are violated. There was some evidence of non-linearity, particularly involving Attention Switching; however, for the most part a linear trend provided a reasonable description of the data.

Table 2.2: Descriptive statistics for ASD, non-ASD and combined samples for the 5 AQ domains

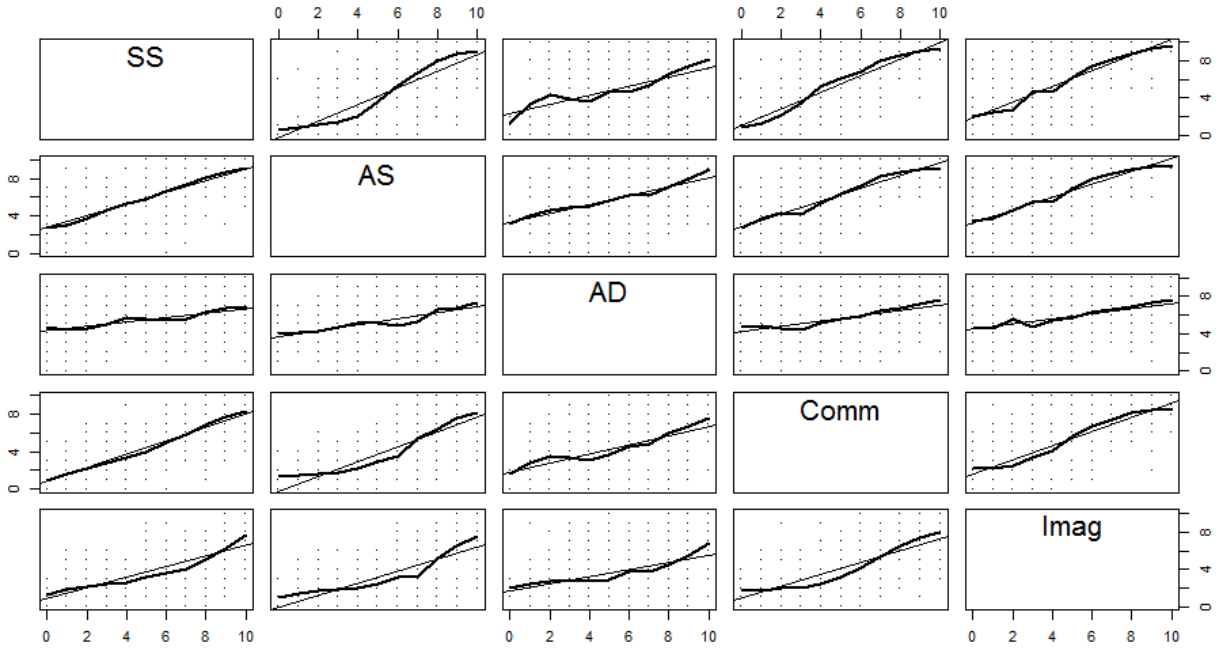
Domain	Mean (SD) Cases	Mean (SD) Control Sample	Mean (SD) Combined sample	$\frac{SD_{Cases}}{SD_{Combined}}$
Social Skills	7.95 (2.14)	3.43 (2.93)	5.14 (3.45)	.62
Attention Switching	8.47 (1.52)	4.47 (2.43)	6.01 (2.88)	.52
Attention to Detail	6.36 (2.30)	5.06 (2.32)	5.58 (2.38)	.97
Communication	7.19 (2.19)	2.94 (2.42)	4.57 (3.10)	.71
Imagination	6.16 (2.51)	2.44 (1.93)	3.82 (2.80)	.90

Figure 2.3: Histograms of AQ subscale scores in combined sample



Cronbach's alpha values for the five domains were: Social Skills = .88, Attention Switching = .81, Attention to Detail = .66, Communication = .83 and Imagination = .78 estimated in the whole sample. Based on the ASD sample alone Cronbach's alpha levels were, as expected, generally lower: Social Skills = .75, Attention Switching = .50, Attention to Detail = .65, Communication = .63 and Imagination = .71.

Figure 2.4: Scatterplot matrix of AQ subscale scores in combined sample



Note. SS=Social Skills, AS= Attention Switching, AD= Attention to Detail, Comm= Communication, Imag=Imagination.

The correlations among symptom domains measured by the AQ in both the combined case and control sample and the case sub-sample are provided in Table 2.3. In the combined sample the correlations between symptom domains ranged from r = .33 to r = .80. With the exception of the Attention to Detail domain which did not correlate well with other symptoms, all of the symptom correlations were large >.65. In the case sub-sample, all of the symptom inter-correlations were substantially smaller than in the combined sample. In the case sub-sample, symptom inter-correlations ranged from r = .21 to .55. The percentage differences between the combined and case sub-sample ranged from -16 to -45%. Therefore, the real data analysis supported the hypothesis that samples restricted to clinically diagnosed individuals could substantially under-estimate symptom inter-correlations.

Range restriction corrected correlations are also provided in Table 2.3. Compared with the uncorrected correlations, these were generally closer to those in the combined sample and were in some cases identical or near-identical to those in the combined sample. The largest percentage differences between the corrected and combined sample estimates involved correlations with the Attention to Detail domain (-33% with Imagination and +28%) with Attention Switching.

Table 2.3: Correlation matrix of the 5 AQ domains in ASD versus combined sample

	SS	AS	AD	Comm	Imag
Combined					
SS	-				
AS	.75	-			
AD	.34	.40	-		
Com	.80	.75	.38	-	
Imag	.68	.66	.33	.69	-
Controls					
SS	-	-20	-35	-13	-29
AS	.60	-	-23	-25	-35
AD	.22	.31	-	-32	-36
Com	.70	.56	.26	-	-35
Imag	.48	.43	.21	.45	-
Cases					
SS	-	-45	-29	-31	-22
AS	.41	-	-25	-32	-32
AD	.24	.30	-	-16	-36
Com	.55	.51	.32	-	-26
Imag	.53	.45	.21	.51	-
Range restriction corrected					
	SS	**AS**	**AD**	**Comm**	**Imag**
SS	-	-24	+9	-9	+4
AS	.57	-	+28	0	+5
AD	.37	.51	-	-13	-33
Com	.73	.75	.33	-	-7
Imag	.71	.69	.22	.64	-

Note: Correlations below the diagonal, % difference relative to combined sample above the diagonal. SS=Social Skills, AS= Attention Switching, AD= Attention to Detail, Com= Communication, Imag=Imagination.

Discussion

In this chapter, I demonstrated that the selection process entailed in diagnosing individuals with ASD may lead to substantial attenuations of symptom inter-correlations in clinically ascertained research samples. I presented evidence that, considering individuals with and without ASD together, the correlations among symptom domains can be quite large, even when only modest in individuals with a clinical diagnosis of ASD. This has implications for the hypothesis that ASD comprises relatively distinct symptoms because it suggests that previous studies utilising clinical samples could have under-estimated the extent to which ASD symptoms correlate with one another.

I used a brief simulation study to estimate the magnitudes of observed correlations between ASD symptoms that could be expected, given different levels of population inter-correlations and a selection mechanism corresponding to ASD diagnosis. I included a range of plausible simulation conditions in terms of the population correlations between symptoms. Consistent with the claim that these represent realistic conditions, all produced ASD prevalence estimates in the range observed in empirical studies which vary by country and methodology but are generally in the range of around 0.2 to 1.6% (Fombonne, 2003; Zaroff & Uhm, 2012). Results showed that symptom inter-correlations are potentially substantially reduced in samples restricted to individuals who meet diagnostic criteria for ASD. Further support for this was found in a real data example in which I compared inter-correlations in an ASD and a combined ASD and control sample. In every case the correlation in the combined sample was substantially larger than in the ASD sub-sample.

The simulation study focussed on the classical triad of ASD because it is within this framework that a large amount of the work on assessing the degree of fractionation of ASD symptoms has been conducted. Similar considerations nonetheless apply to other frameworks or features of ASD such as 'the dyad' of ASD (i.e. the social communication and restricted repetitive activities criteria of DSM 5) or performance on theory of mind or executive function tasks. The extent to which the inter-correlations among these ASD features in clinically diagnosed samples are downwardly biased will depend on several factors. First, it will depend on the population correlation between the features of interest. Under most realistic conditions, the larger the population correlation, the smaller the attenuation of their association in a selected sample (Taylor, 2004). This is because as the population correlation gets smaller, both the numerator and the denominator of the ratio forming the correlation coefficient decrease. However, the numerator (the covariance between the two variables) decreases slower than the denominator (the product of their standard deviations) with

decreasing population correlations. Thus, the percentage bias in small population correlation coefficients will be bigger than that for large population correlation coefficients for a given degree of range restriction.

In the case of the classical triad, this could have accentuated the difference in empirical inter-correlations between the Soc and Comm domains on the one hand and the RSB domain on the other that led to the former two symptoms being combined into a single domain in DSM 5 criteria while the latter remains distinct (e.g. Frazier et al., 2012).

Second, it will depend on how closely the symptoms of interest correspond to the variables on which diagnosis has been made. In range restriction terminology, this is the degree of association between the selection variables and the incidental variables. Stronger associations will result in larger biases. For example, for individuals diagnosed on the basis of DSM-IV, the correlations among the triad should be most strongly affected, with other features of ASD less directly selected affected to a lesser extent.

Third, it will depend on how strong the selection is. While it is the goal of clinicians to diagnose all individuals who genuinely meet the criteria for ASD and none who do not, uncertainty surrounding diagnosis is inevitable. Misdiagnosis of individuals can potentially mitigate the effects of selection on symptom inter-correlations by weakening the strength of selection on ASD symptoms in a clinically diagnosed sample. This can be thought of in terms of the sensitivity and specificity of the instruments used to make a clinical diagnosis and the diagnostic process as a whole. All else being equal, diagnostic procedures with high specificity (usually associated with lower sensitivity) will be associated with stronger selection. Similarly, if diagnostic criteria loosen, resulting in more individuals receiving diagnosis of ASD and corresponding increases in prevalence, then the effect of diagnosis on symptom inter-correlations in such samples will be reduced. It is expected that the move from

DSM-IV to DSM 5 diagnostic criteria may result in a slight reduction in the prevalence of ASD (Maenner et al., 2014); therefore, the effect of diagnosis on symptom correlations may increase in the future as the stricter criteria are adopted. Strength of selection is also affected by the fact that a minority of individuals with a diagnosis of ASD receive that diagnosis in spite of not meeting all diagnostic criteria. To the extent that these individuals are included in empirical studies of clinical samples, this can also mitigate the effects of clinical diagnosis on symptom inter-correlations, by weakening the strength of selection on ASD symptoms.

Finally, though I have framed my demonstration in terms of symptom *inter-correlations* because they have acquired a level of theoretical importance in the literature, the consequences of selection on ASD traits are also not limited to this parameter. Other statistics that depend on the variance or inter-correlation of variables in the sample will also be affected. This includes, for example, the reliability of psychometric assessments (Fife et al., 2012), genetic and environmental variances and correlations (Dominicus et al., 2006), and factor model parameters (Muthén, 1990). For example, the real data example illustrated the attenuation of Cronbach's alpha in clinically diagnosed individuals relative to that of a combined sample of clinically diagnosed and control cases.

Collectively, these considerations suggest that investigations of ASD symptom inter-correlations and related statistics should recruit and jointly analyse data from participants both with and without ASD (e.g. Constantino et al., 2004). This approach is justified if it is assumed that clinical ASD is merely the extreme end of a single trait or confluence of traits that are continuously distributed in the population. That is, autistic traits must be considered meaningful in the general population and not qualitatively different from the traits expressed by individuals with clinical diagnosis of ASD. Such a viewpoint is becoming increasingly accepted. However, this approach also requires resolution of the practical issue of reliably and equivalently measuring ASD traits across both clinically diagnosed and community

populations (Murray et al., 2014). As Happé & Ronald (2008) noted, measures of ASD have face validity in clinically diagnosed samples but when the same measures are administered to individuals without a clinical diagnosis of ASD, it is not clear how the resulting data relates to clinical ASD. Measures of the broader autism phenotype (BAP) or autistic-like traits (ALTs) that aim to capture sub-clinical variation in ASD traits may be advantageous in this regard because they explicitly aim to capture levels of ASD traits that span normality and clinical ranges of the traits (see Wheelright et al., 2010).

Where restricted samples are used, it may be worthwhile estimating range restriction corrected associations. While it is unlikely that they will yield accurate estimates of the 'true' association because this would rely on possibly unrealistic assumptions and accurate estimates of population variances of the variables involved, they can at least provide a sensitivity analysis. In the current study the range restriction corrected estimates in the real data example were, though imperfect, closer to the empirical estimates in the combined sample than were the empirical estimates in the case sample.

Finally, if – as the current results suggest – the estimates of the association between different symptom domains are under-estimated in clinical samples, this suggests that the evidence for a fractionation between ASD symptoms may not be as strong as previously thought. Although it is generally accepted that the social and communication symptoms of ASD have a strong tendency to co-occur (hence their lumping together in DSM 5); there remains much debate about the extent to which these reflect the same syndrome as the non-social symptoms of ASD. Perceptions of the extents of correlation between symptom domains can have a strong influence on psychopathology taxonomies, therefore, it is important to ensure that estimates of these are accurate. However, it is equally important not to rely too heavily on symptom inter-correlations as a basis for nosology. Ultimately, whether conceptualising two symptoms as belonging to the same disorder or not should be based on

whether doing so leads to useful ways of identifying individuals who may have a shared etiology or who may benefit from similar kinds of treatments. Furthermore, it is important to bear in mind that high symptom correlations need not mean a single unitary underlying syndrome: it can also reflect sets of shared causes, local interactions between symptoms, or a set of common end points from a range of causal pathways (i.e. equifinality). Thus, although it is important to obtain accurate estimates of symptom inter-correlations for the purposes of guiding theory and nosology, these alone cannot be relied upon to understand the causal processes underlying their association.

Limitations

The simulation study was designed to reflect the process of diagnosis of ASD; however, because this selection cannot be characterised exactly, the possibility remains that the simulated process did not accurately reflect selection processes in the real world in some way. Second, the study focussed exclusively on symptom inter-correlations which have historically been important in the development of the fractionable triad hypothesis; however, these should not form the sole basis of substantive theory. For example, the fractionable hypothesis has also been informed by conceptual analyses, genetic etiology and neural substrates of the cognitive features of ASD (e.g. Happé & Ronald, 2008). Finally, the real data example was based on a convenience sample which was, therefore, not population-representative. Although the population distribution of AQ scores is not known, it is likely that individuals with high scores were over-represented because of the large number of clinical cases included in the combined sample. The clinical sample may also have been subject to the kind of diagnostic biases discussed in the introduction such as sex differential selection and Berkson's bias. In addition, it is likely that the control samples had been subject to self-selection biases. For example, individuals who perceive themselves to be high in ASD traits or who have relatives with ASD may have a particular interest in participating in an

ASD study. One possibility is that the correlations in the combined sample over-estimated the population correlation due to range enhancement. Range enhancement occurs when scores in the middle of a distribution are under-represented, leading to an over-estimation of an association. Therefore, it is important to note that the point of the example was not to find the 'true' association between the AQ domains, but to demonstrate that the range of trait levels included can have important effects on statistical results.

Conclusions

Samples restricted to individuals who meet the diagnostic criteria for ASD are likely to produce substantial under-estimates of the associations among different symptoms of ASD as a result of range restriction. Given that substantive theories of ASD and the development of diagnostic and treatment processes may depend on the strengths of inter-correlation among features of ASD, it is important to take into account that observed associations in clinically diagnosed groups may not reflect the associations among these features in the population.

Chapter 3: How construct truncation on achievement may have distorted our understanding of the relation between conscientiousness and ability

The use of clinically ascertained samples discussed in the previous chapter is a common and obvious example of how range restricting selection can distort theoretical conclusions. However, construct truncation can operate in far more subtle and hard-to-detect ways. In this chapter I discuss the possibility that construct truncation on achievement has led to spurious evidence for what I label as the intelligence compensation hypothesis (ICH).

Intelligence Compensation Hypothesis

The relations and interactions between personality traits and intelligence have, historically, been of considerable interest in individual differences research (e.g. Ackerman & Heggested, 1997; Austin et al. 2000; Murray, Booth & Molenaar, 2015). A recently emerged hypothesis regarding personality-intelligence interplay is the 'intelligence compensation hypothesis' (ICH). The hypothesis holds that individuals of low cognitive ability become more conscientious in striving to keep their achievement levels on a par with those of their high cognitive ability peers. Individuals high in cognitive ability, on the other hand, can accomplish the same or more with less effort and so have no need to maintain particularly high conscientiousness. In fact, in being able to rely on their cognitive ability, some may allow their levels of conscientiousness to slide. Given the link between both cognitive ability and conscientiousness-related traits and health, academic and occupational outcomes (e.g. Murray & Booth, 2015; Poropat, 2009; Wrulich, Brunner, Stadler, Keller & Martin, 2014), the ICH could have important practical as well as theoretical implications. Confirming and characterising a possible antagonistic relationship between cognitive ability and conscientiousness could facilitate a more nuanced understanding of the interplay between risk

factors for maladaptive behaviours and associated implications for prevention and intervention.

The ICH grew out of the observation that, in many samples, negative associations between cognitive ability and conscientiousness-related personality traits have been observed (e.g. Furnham, Moutafi, & Chamorro-Premuzic, 2005; Furnham et al., 2007; Moutafi et al., 2003; Moutafi, Furnham, & Paltiel, 2004; Moutafi, Furnham, & Crump, 2006; Furnham & Moutafi, 2012; Soubelet & Salthouse, 2011; Wood & Englert, 2009).

In spite of the intuitive appeal of the ICH, evidence for the hypothesis is mixed. Counter to the hypothesis, some positive associations between cognitive ability and conscientiousness have been observed (e.g. Baker & Bishel, 2006; Lounsbery et al., 2005; Luciano et al., 2006) and other studies have yielded associations that were close to zero or non-significant (e.g. Bartels et al., 2012; Chamorro-Premuzic et al., 2005; Furnham et al., 2005).

The role of sample composition

A feature which appears to distinguish studies supporting the ICH is sample composition. Specifically, the majority of these studies have been conducted in samples which appear to have been subject to selection with respect to occupational or academic achievement. The studies of Moutafi et al. (2004) and Furnham and Moutafi (2012) used samples of junior to middle managers attending staff development centres, whilst other studies have utilised samples of managerial grade job applicants attending assessment centres (Furnham et al., 2007; Wood & Englert, 2009). Development and assessment centres are costly (Eurich, Krause, Cigularow, & Thorton, 2000). This means that organisations tend to invite only a small percentage of the total applicant pool when the purpose of attendance is to provide information for making selection decisions, and in training contexts, it is usually

individuals from managerial and professional populations who attend (Meriac, Hoffman, Woehr, & Flisher, 2008; Pepermans, Vloeberghs, & Perkisas, 2003). Similarly, the study by Furnham et al. (2005) used a sample of undergraduate students and entry to university involves selection on prior academic achievement (e.g. Hägglund & Larsson, 2006). The study by Soubelet & Salthouse (2011) used a sample recruited via newspaper adverts and referrals from other participants. Although the sample was not selected in such an obvious way as a professional or student sample might be, the recruitment process and exclusion criteria (participants with Mini-Mental State Examination scores that indicated potential cognitive impairment were not eligible to participate; Folstein, Folstein, & McHugh, 1975) resulted in a sample who had an average of almost 16 years of education and were approximately 2/3 to 1 standard deviation above the national norms on cognitive ability.

Compensatory selection

The selected compositions of these samples raises the possibility that the apparent negative associations between intelligence and conscientiousness-related traits were due not to individual calibration of conscientiousness levels to ability level as proposed by the ICH, but to compensatory *selection* into the populations from which the research samples investigating the question are taken (see Sackett, Lievens, Berry, & Landers, 2007). To enter the population of individuals employed in professional jobs or the population of individuals undertaking university level education, a certain level of achievement (educational or occupational) is necessary. Compensatory selection refers to a process whereby selection into these populations through meeting these achievement criteria can be done through combinations of ability and hard work (i.e. conscientiousness). An individual of low conscientiousness can still enter the population if they are of high ability and an individual of low ability can still enter the population if they are of high conscientiousness; however, individuals with a combination of low conscientiousness and low ability are excluded.

Operationalising this situation statistically, one could think of selection into the research sample being based on a composite of IQ and conscientiousness:

$$Ach = b_1 IQ + b_2 con, \tag{3.1}$$

where b_1 and b_2 are weights determining the contributions of IQ and conscientiousness to achievement. From eq. 3.1 it is obvious that whenever IQ is relatively low, high composite values can still be observed so long as high conscientiousness compensates for it. Setting an appropriate 'achievement threshold', individuals with co-occurring low conscientiousness and IQ will certainly be excluded; however, those with low values on one trait but high on the other may exceed the threshold. A research sample based on a population selected in this way could yield a negative correlation between IQ and conscientiousness even if they are not correlated or even positively correlated in the population because it will tend to have a greater proportion of people with discrepant IQ-conscientiousness scores than the general population.

The conceptual description above can be considered in more formal terms. Sackett et al. (2007) outlined the situations under which truncation of a composite such as that in Eq. 3.1 would lead to spurious negative associations between its constituents. They noted that the correlation between components of the composite is attenuated according to the following equation, in which, for this chapter, I have presented in terms of the variables 'ach', 'IQ' and 'con', standing for 'achievement', 'intelligence quotient (i.e. cognitive ability)' and 'conscientiousness' respectively:

$$r'_{IQ,con} = \frac{r_{IQ,Con} - r_{IQ,ach}r_{con,ach} + r_{IQ,ach}r_{con,ach}(\frac{V'_{ach}}{V_{ach}})}{\sqrt{1 - r^2{}_{IQ,ach} + r^2{}_{IQ,ach}(\frac{V'_{ach}}{V_{ach}})}\sqrt{1 - r^2{}_{con,ach} + r^2{}_{Con,Ach}(\frac{V'_{ach}}{V_{cch}})}}$$

(3.2)

where $r'_{IQ,Con}$ is the correlation between IQ and conscientiousness in the sample truncated with respect to achievement, $r_{IQ,con}$ is the population correlation between IQ and conscientiousness, $r_{IQ,Ach}$ is the population correlation between IQ and achievement, $r_{con,Ach}$ is the population correlation between conscientiousness and achievement, V'_{Ach} is the variance of achievement in the truncated sample and V_{Ach} is its variance in the population. The correlations between IQ and conscientiousness depend on the regression weights in Eq. 3.1. Truncation on Achievement can produce a negative association between Conscientiousness and IQ in a sample even when they are positively correlated in the population if the product of the correlations of each with Achievement (the second term of the numerator) is greater than their inter-correlation (the first term of the numerator). In these situations, subtracting this second term from the first will result in a negative number. This negative number is then added to the third term which is the product of the correlations between conscientiousness and Achievement, IQ and Achievement and the selection ratio capturing the extent of range restriction on Achievement. If the selection is a very small number, reflecting strong selection, the third term of the numerator as a whole will be small and fail to offset the negative number yielded by subtracting the second from first term. Here, a negative correlation between conscientiousness and ability will arise in the sample. For this to happen, conscientiousness and IQ would need to be relatively strongly correlated with Achievement but relatively independent of one another because under these circumstances the value of second term of the numerator would exceed the value of the first term. From a theoretical perspective, this seems plausible, suggesting that the situation

identified by Sackett et al. (2007) could apply to the conscientiousness-IQ associations cited as evidence of ICH.

ICH, compensatory selection and achievement-IQ interactions

To summarise the distinction between ICH and compensatory selection, the processes implied by the ICH suggest that there is a causal impact of ability on conscientiousness: a calibration of conscientiousness levels to ability. Compensatory *selection* invokes no such causal effect - it merely applies to situations where reaching any threshold required for selection into a sample can be accomplished through many different combinations of conceptually different variables, even when the threshold is only indirectly stated or assessed. In the ICH a negative association between ability and IQ is predicted in the population. On the other hand, if compensatory selection is true, it is more likely to be zero or at most very weakly positive or negative.

Chapter aims

In this chapter, I tested the hypothesis that the negative IQ-conscientiousness association observed in many previous studies is an artefact of truncation on achievement. To do so I used a sample for which there was little evidence of selection on educational and occupational achievement and which could, therefore, be considered reasonably free of achievement truncation. I also assessed the extent to which a negative association between conscientiousness and IQ could be induced by artificially introducing truncation on educational or occupational achievement. The purpose of this was to simulate the processes that may have occurred during the selection of many of the samples used to evaluate the conscientiousness-IQ relation. I also tested this compensatory selection hypothesis against a moderated association hypothesis in order to assess whether any apparent effects of compensatory selection simply reflected moderation of the effect of IQ on conscientiousness

by achievement. I hypothesised that 1) I would not find significant negative associations between conscientiousness traits and IQ in the whole (untruncated) samples, 2) negative associations could be induced by selection on educational achievement (in an adolescent sample) and occupational achievement (in a parent sample).

Method

Participants

I analysed data from the Minnesota Twin Family Study (MTFS) and Sibling Interaction and Behaviour Study (SIBS). MTFS is a community-based longitudinal study of same-sex twins and their parents. Participants were recruited using a population-based method (for a full description see Iacono, Carlson, Taylor, Elkins, & McGue, 1999). It consists of two cohorts, one recruited when the twins were aged 11 years, and the other when the twins were aged 17. Both cohorts have been followed approximately every three years since this initial wave of data collection. Compared to the US Census data for Minnesota, the MTFS sample appears to be generally representative of families with children living at home (Holdcraft & Iacono, 2004). Approximately 20% of invited participants declined to participate; however, more than 80% of those who declined completed a brief mail or telephone survey. This allowed a partial comparison of individuals who agreed to participate with those who did not so that any important differences could be identified. Parents in participating families had on average an additional 0.3 years of education (for additional comparisons see Iacono et al, 1999), which was judged to be only a small difference unlikely to have any practical influence on the current study.

SIBS is a community-based sample of pairs of adoptive and biological siblings and their parents recruited through adoption agencies. The families comprising the adoptive sample were selected to include an adolescent between the ages of 10 and 21 who was

adopted before the age of 2 and a second adolescent who was not biologically related and was no more than five years older or younger. The parents in these families were broadly representative of those accepting infant placements, but compared with Minnesota parents in the general population they were of higher socioeconomic status. The families in the biological families were recruited using birth records from the same area as the adoptive families. Fifty-seven percent of eligible biological families agreed to participate and 63% of eligible adoptive families agreed to participate; however, 90% of the mothers from the remaining families completed a brief telephone interview, again allowing comparison of those who agreed and declined to participate. These groups did not differ on either educational or occupational level among the adoptive families but mothers from the participating biological families were more likely to have a college degree than those from non-participating families (44% of the sample had a college degree compared with an 39% for the comparison population of mothers in the same geographical region).

Overall, therefore, the combined sample was slightly selected on parental education and socio-economic status but otherwise generally representative of individuals in the geographic region from which they were sampled and of parents of adoptive children.

Adolescent Sample.

In the adolescent sample, I used data from the 11- and 17-year-old MTFS cohorts and SIBS. I combined the data from the second wave of follow up in the 11-year-old cohort (targeting them at age 17) with the intake data from the other cohorts. Dependent on the data available on particular measures, I used different subsets of the total sample. The composition of these samples varied slightly but as an approximate guide, with complete data on the IQ and both measures of conscientiousness, there were 2412 participants (1100 males) with a mean age of 17.7 (SD = 0.69)

Parent Sample.

I combined the parent data from the MTFS and SIBS cohorts, utilising data contributed at intake. Again, the specific subset of data used from the sample as a whole was dependent on the availability of particular measures. As an approximate guide, with complete data on the IQ and both of the conscientiousness measures, there were 3276 participants (1522 males) with a mean age of 42.5 (SD = 5.5).

Measures

Multidimensional Personality Questionnaire (MPQ).

Conscientiousness facets were measured using a 198-item version of the multidimensional personality questionnaire (MPQ; Tellegen & Waller, 2008). The MPQ contains two conscientiousness-related traits: Control and Achievement. For this chapter, I re-label the Achievement scale 'Achievement-striving' to avoid confusion with the measures of occupational and educational achievement. High scorers on Control endorse being reflective; cautious, careful, plodding; rational, sensible, level-headed, liking to plan activities in detail. High scorers on Achievement-striving endorse being hard-working, driving themselves, welcoming difficult and demanding tasks; persisting when others give up; ambitious , putting work and accomplishments before many other things setting high standards and being perfectionistic. Items were measured on a 4-point response scale from 'Definitely True' to 'Definitely False' and each scale had 18 items. Here I utilised the summed scale scores for the two measures. I analysed the facets of Control and Achievement-striving separately because there is growing evidence that facets within the domain of Conscientiousness have differential criterion and outcome associations (e.g. Bogg & Roberts, 2013). Thus analysing associations at the level of the broader dimension of conscientiousness risks obscuring substantively important facet-level processes.

Discussing their relations to the five-factor model, Gaugham, Miller, Pryor, and Lynam (2009) reported the highest correlations of MPQ Control to be with the Order (r = .56) and Deliberation (r = .68) facets of Conscientiousness in the NEO-PI-R, whilst MPQ Achievement-striving correlated most highly with the Achievement Striving (r = .60) and Self-Discipline (r = .52) facets.

IQ

The IQ measure completed by participants was an abbreviated version of the Wechsler Adult Intelligence Scale Revised (WAIS-R; Wechsler, 1974) and included the Vocabulary, Information, Block Design and Picture Arrangement subtests. These subtests were chosen based on their high correlation (r =.90) with full scale IQ derived from all the subtests.

Educational and Occupational Achievement.

For the adolescent sample I used grade point average (GPA) as a measure of educational achievement. To avoid problems of comparing grades across different school districts with different testing formats, procedures and standards, GPA was not computed from actual grades. Instead twins and their parents were asked to report, on a 5-point scale from 0 = failed class to 4 = much better than average, the grades typically received in language arts, maths, social studies and science classes. Here, GPA was the average across these ratings. This measure was validated against the actual school grades of a sub-sample of 67 randomly selected participants from the age-11 cohort and found to correlate with these at .89.

For the parent sample, I used occupational level according to the Hollingshead's (1957) occupational scale as a measure of occupational achievement. This is an eight-point

scale ranging from ‘unskilled’ to ‘major professional’. Higher ratings on the scale reflect higher levels of occupational achievement.

Statistical Procedure

Compensatory Selection

I chose methods of evaluating the correlation between IQ and our two measures of conscientiousness (Control and Achievement-striving) were designed to mimic as closely as possible the methods that have been employed in the majority of previous studies finding negative associations between IQ and conscientiousness (e.g. Moutafi et al., 2004). I, therefore, used Pearson’s correlations between the scale scores on the personality measures and IQ. I dealt with missing data using pairwise deletion.

I introduced truncation on achievement by discarding all individuals who were below progressively stricter thresholds of educational or occupational achievement. This was designed to mimic processes of selection into populations (e.g. undergraduate students, or assessment centre participants) in a manner that was dependent on educational or occupational achievement. I then evaluated the correlations between IQ and the conscientiousness measures in each of the progressively more selected samples.

Evaluation of IRR formula

The formula in Eq. 3.2 is a special case of Thorndike case III i.e. indirect selection. To evaluate whether the estimated correlations based on this formula tally with those in empirical data subject to selection, I applied the formula in Eq. 3.2 to the data at each level of selection. A lack of correspondence between the estimated and actual correlations would suggest that Eq. 3.2 has limited utility in predicting (and conversely, correcting for) the effects of selection in practice.

Moderation analysis

It was not possible to directly compare a moderation model with a model of compensatory selection directly in the sense of estimating both and comparing their fits. However, the absence of a moderation effect would suggest that any apparent compensatory selection effect was not really just due to an interaction between achievement and IQ. I, therefore, also evaluated whether educational or occupational achievement moderated the effect of IQ on conscientiousness. To do this I used moderated multiple regression models. One model was estimated for each of the measures of conscientiousness in each of the samples. In these models the predictors were IQ, achievement (occupational level for the parent sample and GPA for the adolescent sample) and their product. The outcome variable was the conscientiousness measure (Control or Achievement-striving). IQ and achievement were both centred prior to analysis. A statistically significant interaction term was considered to be evidence in favour of moderation of the relation between IQ and conscientiousness by achievement.

Results

Correlations in unselected samples

In the unselected adolescent sample there was no statistically significant association between IQ and Control (r = .04, p = .06) but a statistically significant positive association between IQ and Achievement-striving (r = .14, p<.01). In the unselected parent sample there was a small but statistically significant positive association between IQ and Control (r = .05, p<.01) but no statistically significant association between IQ and Achievement-striving (r = .03, p = .15).

Effect of selection on conscientiousness-IQ association

Tables 3.1 and 3.2 show the correlations of IQ with the Control and Achievement-striving personality scales when the full samples were subjected to selection on educational or occupational achievement. They show the downward trajectories of the correlations as samples became increasingly selected on achievement or occupational achievement. This is depicted graphically in Figures 3.1 and 3.2. The error bars indicate 95% confidence intervals.

Application of Thorndike case III

In the adolescent participants, the whole sample correlations between IQ and GPA, Control and GPA, and Achievement-striving and GPA were r=.38, r=.29 and r=.32 respectively. The whole sample and selected sample variances are provided in Table 3.3 together with the predicted correlations based on Thorndike case III formula. The predicted correlations were in most cases quite close to the actual correlations. Using Thorndike case III, a negative association between Control and IQ but not between Achievement-striving and IQ would have been predicted at the highest levels of selection.

In the parent participants, the whole sample correlations between IQ and occupational achievement, Control and occupational achievement, and Achievement-striving and occupational achievement were r=.40, r=.10 and r=.11 respectively. The whole sample and selected sample variances are provided in Table 3.4 together with the predicted correlations based on Thorndike case III formula. The predicted and actual correlations were for the most part similar, diverging only for the IQ-Achievement association and only at the highest level of selection. At very high levels of occupational achievement, the range restriction formula suggested only a very small negative correlation between Achievement-striving and IQ, however, a much larger negative correlation was observed empirically. This suggests that the association between occupational level and the combination of IQ and Achievement-striving

that facilitates occupational success is non-linear, possibly even substantively different at the highest levels of occupational achievement.

Table 3.1: Correlations between IQ and conscientiousness at different levels of selectivity for educational achievement in adolescent sample

Selection criterion	**IQ-Control correlation**			**IQ-Achievement-striving correlation**		
	R	***N***	***P***	***r***	***N***	***p***
No selection	.04	2416	.06	.14	2417	<.01
GPA>1	.03	2285	.09	.15	2285	<.01
GPA>1.25	.03	2270	.11	.15	2270	<.01
GPA>1.5	.03	2240	.14	.15	2239	<.01
GPA>1.75	.03	2196	.20	.14	2195	<.01
GPA>2	.02	2066	.47	.13	2065	<.01
GPA>2.25	.00	1964	.92	.13	1963	<.01
GPA>2.5	-.02	1758	.38	.12	1758	<.01
GPA>2.75	-.02	1538	.48	.12	1539	<.01
GPA>3	-.04	1236	.18	.10	1237	<.01
GPA>3.25	-.05	1014	.08	.09	1014	<.01
GPA>3.5	-.07	662	.07	.10	663	.01
GPA>3.75	-.06	375	.22	.08	376	.13

Table 3.2: Correlations between IQ and conscientiousness at different levels of selectivity for occupational achievement in adult sample

Selection criterion	IQ-Control correlation			IQ-Achievement-striving correlation		
	r	***N***	***P***	***r***	***N***	***p***
0. No selection	.05	3280	<.01	.03	3277	.15
1. Semi-skilled and above	.04	2332	.04	.01	2329	.61
2. Skilled manual and above	.03	2247	.15	.00	2090	.96
3. Clerical, sales, technician etc. and above	.02	1696	.50	.01	1694	.74
4. Minor professional and above	.01	1293	.69	-.05	1291	.10
5. Lesser professional and above	.01	743	.79	-.05	742	.16
6. Major professional and above	.03	269	.64	-.13	268	.03

Figure 3.1: Conscientiousness-IQ associations in adolescent sample

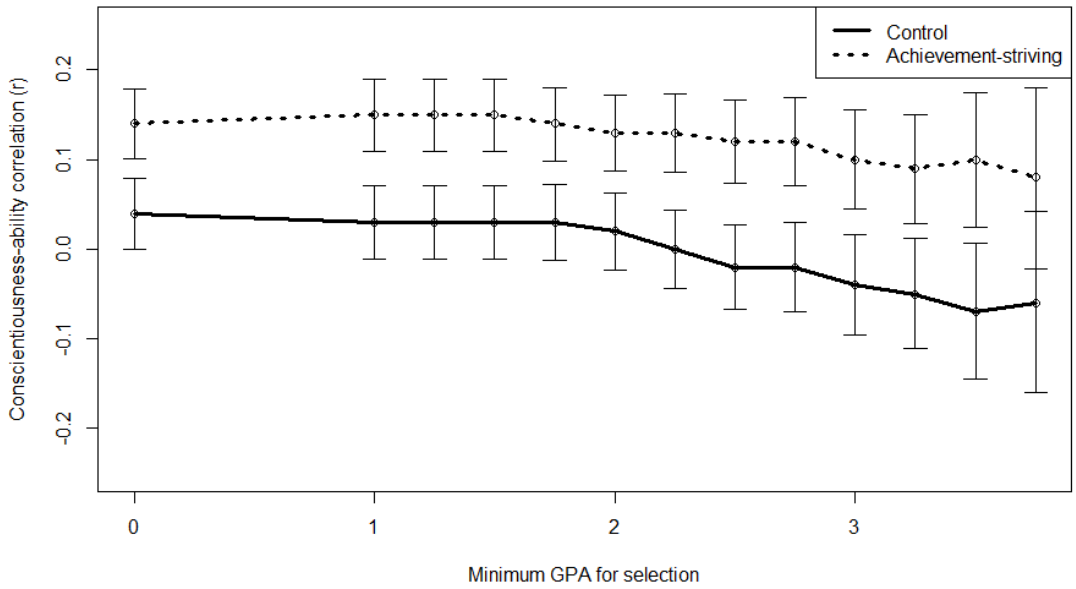


Figure 3.2: Conscientiousness-IQ associations in parent sample

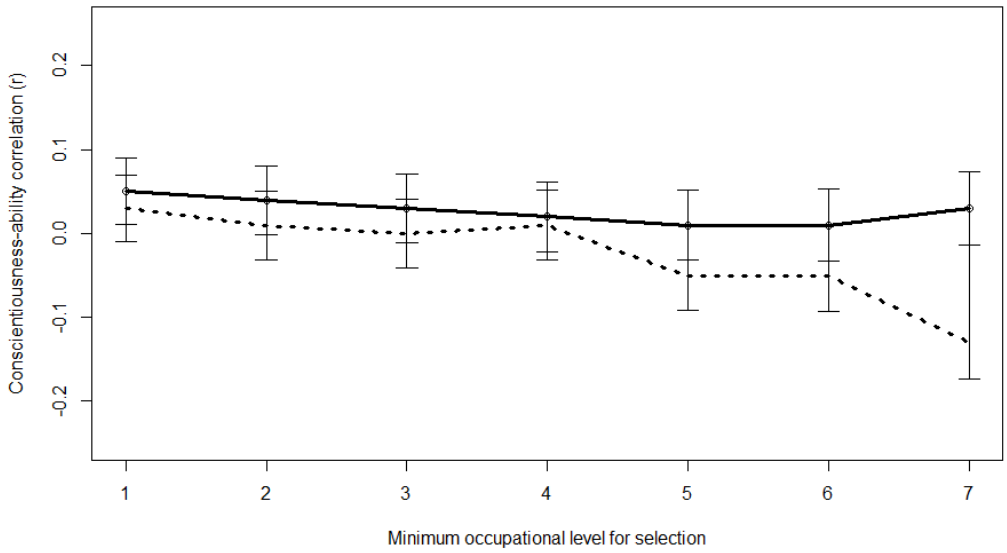


Table 3.3: Application of Thorndike case III to adolescent sample correlations

Selection criterion	Standard deviations				IQ-Control correlation			IQ-Achievement-striving correlation		
	GPA	**IQ**	**Control**	**Ach**	***r***	***r′***	***r-r'***	***r***	***r′***	***r-r'***
No selection	.72	14.53	7.60	8.49	.04	.04	.00	.14	.14	.00
GPA>1	.67	14.50	7.52	8.40	.03	.03	.00	.15	.13	.02
GPA>1.25	.66	14.50	7.52	8.39	.03	.02	.01	.15	.12	.03
GPA>1.5	.63	14.47	7.50	8.36	.03	.01	.02	.15	.11	.04
GPA>1.75	.60	14.46	7.52	8.35	.03	.01	.02	.14	.11	.03
GPA>2	.54	14.45	7.43	8.29	.02	-.01	.03	.13	.09	.04
GPA>2.25	.50	14.45	7.45	8.22	.00	-.02	.02	.13	.08	.05
GPA>2.5	.43	14.49	7.45	8.23	-.02	-.03	.01	.12	.07	.05
GPA>2.75	.36	14.45	7.42	8.18	-.02	-.05	.03	.12	.05	.07
GPA>3	.28	14.43	7.43	8.15	-.04	-.06	.02	.10	.04	.06
GPA>3.25	.23	14.60	7.51	8.24	-.05	-.07	.02	.09	.03	.06
GPA>3.5	.13	14.33	7.42	8.21	-.07	-.07	.00	.10	.03	.07
GPA>3.75	0.00	14.18	7.40	7.85	-.06	-.08	.02	.08	.02	.06

Note. r is empirical estimate, *r*' is estimate based on Thorndike Case III formula. Occ= Occupational achievement, Ach= Achievement-striving.

Table 3.4: Application of Thorndike case III to parent sample correlations

Selection criteria	Standard deviations				IQ-Control correlation			IQ-Achievement-striving correlation		
	Occ	**IQ**	**Control**	**Ach**	***r***	***r'***	***r-r'***	***r***	***r'***	***r-r'***
0.	1.64	14.21	7.19	7.80	.05	.05	.00	.03	.03	.00
1.	1.52	14.22	7.24	7.68	.04	.04	.00	.01	.02	-.01
2.	1.30	14.11	7.24	7.67	.03	.04	-.01	.00	.01	-.01
3.	1.01	14.31	7.31	7.59	.02	.03	-.01	.01	.00	.00
4.	0.77	14.16	7.42	7.46	.01	.02	-.01	-.05	.00	-.05
5.	0.48	13.85	7.50	7.54	.01	.01	.00	-.05	-.01	-.04
6.	0.00	12.95	7.82	7.94	.03	.01	.02	-.13	-.01	-.12

Note. r is empirical estimate, *r*' is estimate based on Thorndike Case III formula. Occ= Occupational achievement, Ach= Achievement-striving. See Table 3.2 for selection criteria.

In the adolescent sample, the initial non-significant positive association between IQ and Control in the full sample ($r = .04$, $p = .06$) became steadily attenuated and then negative with selection on GPA. At the highest level of GPA, the association was $r = -.06$ ($p = .22$). A similar albeit more subtle effect occurred in the correlation between IQ and Achievement-striving, which began at $r = .14$ ($p<.01$) and decreased to $r = .08$ ($p = .13$) in the most selected group.

In the parent sample, selection on occupational level had little effect on the correlation between IQ and Control. It reduced from .05 to .01 and then rose again to .03 at the highest level of selection. There was a more marked effect of selection on the correlation between IQ and Achievement-striving. With increasing degrees of selection, it first became steadily attenuated to zero with and then became negative. Although there was no significant

association between IQ and Achievement-striving in the full sample, at the highest level of selection there was a statistically significant negative association (r = -.13, p = .03).

Moderation tests

There was no statistically significant interaction between IQ and GPA in predicting either Control (b = 0.02, p = .32), or Achievement-striving (b = 0.04, p = .08) in the adolescent sample. There was also no statistically significant interaction between IQ and occupational level in predicting either Control (b = -0.00, p = .86) or Achievement-striving (b = -0.00, p = .62) in the adult sample. These results suggest that achievement did not moderate the effect of IQ on conscientiousness.

Discussion

In this chapter I tested whether compensatory selection into research samples could explain why negative associations have been observed between conscientiousness and cognitive ability. Often these associations are explained in terms of an 'intelligence compensation hypothesis' in which lower ability individuals develop higher levels of conscientiousness to compensate for their lower ability. Many studies have, however, failed to find the expected negative associations between IQ and conscientiousness. Moreover, those that have tended to comprise participants above certain levels of educational or occupational achievement.

I found no evidence for negative correlation in a large sample of adolescents and their parents. Unlike many previous studies, in this sample only relatively trivial selection on educational or occupational achievement was likely. Where there were significant associations between IQ and conscientiousness in the full sample, these were positive rather than negative. In fact, there was a positive correlation between IQ and Achievement-striving

(r =.14) of an absolute magnitude comparable to the negative correlations reported in previous studies cited in support of ICH (e.g. Moutafi et al., 2006).

However, this was not the first study to report evidence contradicting ICH as others have found no significant association or small positive associations have been conscientiousness and IQ (Bartels et al., 2012; Lounsbury et al., 2005; Luciano et al., 2006). Notably, like the current study, many of these studies did not appear to show evidence of substantial sample selection on achievement.

The general pattern of zero to small associations between IQ and conscientiousness in studies apparently not selected on achievement might suggest one of two causal scenarios at the level of the individual. Either there are only minimal causal impacts of IQ and conscientiousness on one another; or the impacts of IQ and conscientiousness on one another are heterogeneous across individuals but close to zero in the aggregate as effects in opposite directions cancel out. For example, while some individuals of lower ability may develop increased conscientiousness in compensation, others of low ability may become discouraged by their failure to achieve on a par with their more able peers without intensified efforts. These latter individuals may grow less conscientious in expending achievement-related effort in response to the lower pay-off they receive for this behaviour. Conversely, the higher rewards for behaving conscientiously in more able individuals could lead to greater reinforcement of this behaviour. A person's particular social environment (e.g. the rewards associated with intelligent and conscientious behaviour) in combination with their other traits (e.g. motivation, reward sensitivity, locus of control, expectations surrounding achievement) will likely also influence whether and how their level of intellectual ability and conscientiousness impact one another.

Soubelet and Salthouse (2011) have suggested that how personality traits and cognition relate to one another may depend on a person's age. Our results support this idea to some degree: only in our adolescent sample was a positive association observed between Achievement-striving and IQ. A possible explanation for this is that adolescents are likely to be currently or recently in academic environments: social settings in which intellectual achievement is heavily measured and rewarded. The salience of intellectual achievement may foster social influences that result in enhancement of conscientiousness particularly in those individuals of higher cognitive ability for whom these rewards are more attainable, with individuals of lower cognitive ability possibly even becoming disheartened and demotivated. Such processes are likely to be governed by a 'frog pond' effect whereby it is not only the absolute level of intellectual ability of individuals that matters with regards to influences on conscientiousness, but also their levels of cognitive ability relative to immediate peers (e.g. see Marsh et al., 2007). Therefore, individuals who perceive their potential for achievement to be more limited because of their relative and absolute cognitive ability would be less likely to strive towards these achievements and thus score lower on conscientiousness.

The primary aim of this chapter was to assess the hypothesis that achievement truncation can account for previously observed negative associations between conscientiousness and ability. Consistent with this, I found evidence that selecting on educational or occupational achievement biased the associations in the negative direction. In the adolescent sample, positive associations between IQ and the conscientiousness measures in the full sample were reduced to negative or effectively zero as subsamples were increasingly restricted to high levels of GPA. In the adult sample there was little effect of restricting the sample to increasingly high levels of occupational achievement on the correlation between IQ and Control. Restricting the sample in this way, however, induced a negative and statistically significant association between IQ and Achievement-striving in

spite of there being no significant association in the full sample. This negative association was of a similar magnitude to those interpreted as evidence for intelligence compensation in previous studies. I also checked whether these apparent compensatory selection effects simply reflected unmodelled moderation of the relation between IQ and conscientiousness by achievement. Moderation effects were very small and non-significant, suggesting that this was not the case. I, therefore, interpreted results as suggesting that achievement truncation may have accounted for some previous observations of a negative association between conscientiousness-related traits and IQ. The fact that only in one out of the 4 cases examined were negative associations induced by selection suggests, however, that at most truncation contributes to, rather than completely explains the previously observed negative associations. Nonetheless, results suggested that differing degrees of selection on achievement could contribute to cross-study differences in the magnitude and direction of association between conscientiousness and IQ.

Unfortunately, it is not possible to ascertain from the study reports the precise selection processes that led participants to be in the research samples in which negative conscientiousness-IQ associations have been observed. For this reason, I cannot be certain that these processes were closely approximated by the simulated selection I used. This is a general problem in observational research: it is uncommon for the selection processes leading to the composition of convenience samples to be explicitly considered, even less to be measured and modelled (see Hunt & Madhyastha, 2008 for a discussion). Unless such selection processes are given due consideration, researchers risk being misled as to the direction and magnitude of the associations between study variables.

Finally, while I have argued in this chapter that variability in sample selectivity on achievement may explain some of the heterogeneity in association between conscientiousness and ability in the published literature, this will not be the only factor influencing the

magnitude of association. For example, different facets of conscientiousness appear to show varying associations with IQ and there may be plausible theoretical interpretations for these differential associations (e.g. Luciano et al., 2006). For example, the 'Competence' facet of Conscientiousness measures may be more positively related to IQ than other facets if it essentially acts as a self-report measure of IQ (e.g. see Chamorro-Premuzic et al., 2005). Similarly, I have argued here that Achievement-striving may be particularly influenced by IQ because motivation to achieve is likely to be influenced by self-perceptions of capacity to achieve. Depending on which facets are measured and whether these are combined into a single Conscientiousness score will, therefore, affect the observed association with IQ.

Finally, the measures of occupational and educational achievement were imperfect. Both scales were coarse and self-reported. Replicating results using alternative sources of information regarding achievement would be valuable. For example, having teacher ratings or school records in the case of GPA would help to ensure that results were not overly influenced by any kind of reporting bias.

Chapter 4: A comparison of alternative phenotypic proxies in tests of gene-environment interactions under construct truncation

The previous chapters have considered construct truncation due to person selection; however, item selection may be an equally important source of construct truncation. One area where construct truncation due to item selection is particularly important is when the construct is hypothesised as the outcome of an interaction. Here, construct truncation can result in significant distortions of estimates of the interactive process, even reversing its apparent direction. In this chapter, I used a simulation study complemented by a real data example to evaluate three possible methods of dealing with this problem in the context of testing gene-environment interactions.

Gene-environment interactions

Increasingly, theoretical perspectives on phenotypic development and expression are recognising that genes and environments transact in dynamic and complicated ways. Many posit some kind of gene-environment interaction (GxE) where GxE is defined as a differential response to environmental circumstances depending on genotype, or, a differential genetic expression depending on environment (Boomsma, & Martin, 2002; Eaves, Last, Marin, & Jinks, 1977). GxE plays a central role in major theoretical models such as the diathesis-stress model, the differential susceptibility model, the vantage sensitivity model, and the bioecological model (Brofenbrenner & Ceci, 1994; Pluess & Belsky, 2013; Reiss, Leve, & Neiderhiser, 2013; Rende & Plomin, 1992). The diathesis-stress model, for example, predicts that the genetic variance in a psychopathological trait is greater in more adverse environments whereas the bioecological model predicts that the genetic potential for a positive trait, such as intellectual ability, is realised to a greater extent in a more stimulating, higher quality environment (Asbury, Wachs, & Plomin, 2005; Rende &

Plomin, 1992). GxEs are also cited as mechanisms by which social factors regulate behaviour, for example, in the theory that genetic influences on certain phenotypes are prevented from being expressed in environments where there are stronger social norms or explicit prohibitions relating to those phenotypes (Shanahan & Hofer, 2005).

Construct truncation in gene-environment interactions

To keep pace with these theoretical developments, it has been necessary to develop statistical methodologies capable of modelling the more complex forms of interplay implied by theory (e.g. Purcell, 2002). Despite the promise and widespread uptake of these approaches, the ability to test theoretically implied GxE interactions is hampered in practice by dependency of tests of interactions on the observed distributions and scales of the phenotype (Eaves et al., 1977, 2002; Eaves, 2006; Mather & Jinks, 1971; Purcell, 2002; Schwabe & van den Berg, 2014).

The problem of dependency of GxE on the scaling of the phenotype has been known since the time of R.A. Fisher noted that GxE interactions could be manipulated by re-scaling the variables involved. In fact, he went so far as to advocate 'transformations of scale' to eliminate what he perceived to be nuisance non-additivity (Tabery, 2008). This suggestion was controversial because he was recommending purging the same non-additivity that was, and still is, viewed by many substantive researchers as a meaningful clue as to the causal processes underlying phenotypic development. Since then, numerous methodological studies have further discussed and demonstrated the dependency of appearance of presence of GxE on scaling (Eaves et al., 1977; Martin, 2000; Molenaar, van der Sluis, Boomsma, & Dolan, 2012; Purcell, 2002; Tucker-Drob, Harden, & Turkheimer, 2009; van der Sluis, Dolan, Neale, Boomsma, & Posthuma, 2006). In the section that follows I summarise and extend the key arguments of these authors.

The key challenge is in the multiplicity of possible causal structures underlying the same sample phenotypic distribution. Consider the case where the observed distribution of the phenotype is non-normal: a common occurrence in behaviour genetic research, as well as psychological research in general (Beasley, Erickson, & Allison, 2009; Miccerri, 1989). The primary problem for testing GxE is that when an observed phenotypic distribution is non-normal, this non-normality could reflect the presence of GxE, or it could simply be that the population distribution of the phenotype is normal but its measurement is poorly scaled so that the observed distribution does not reflect the population distribution. A statistical test of GxE will not be able to distinguish these possibilities.

This difficulty is not one limited to a choice between a 'GxE' explanation and a 'scaling' explanation. There are many biologically plausible alternative models that can produce similar patterns in observed data and, in turn, similar model fits when formally tested. For example, a non-linear main effect of one variable on another is difficult to distinguish statistically from GxE (Rathouz et al., 2008). However, in the current thesis, I focus on scaling specifically because there is considerable evidence that at a large number of phenotypic measures may be vulnerable to the effects of suboptimal scaling.

Cases in point are measures of traits which originated in psychopathological paradigms. These very commonly yield observed non-normal (positively skewed) distributions because majorities of participants score close to the low (non-pathological) ends of the measurement scales. It is often argued that these observed distributions are not necessarily appropriate representations of the population distributions of the phenotypes but arise as a result of the scales being developed with focus on the upper extremes of the traits (van den Oord, Pickles, & Waldman, 2003;van den Oord et al., 2000). Thus, failure to observe a normal distribution for a trait may be a result of failing to measure that trait with

items that have an appropriate range of difficulties to provide reliable coverage of the whole trait distribution.

This position is supported by the observation that even when a case can be made that a clinical or personality trait has a continuous normal distribution in the population, measures of that trait often exhibit item difficulties that are tightly clustered in the impaired range (Meijer & Egberink, 2012; Thomas, 2011). These scales have high discrimination in and around clinical cut-off points but poor discrimination in the healthy ranges. Thus, in a population-representative sample that would include predominantly healthy participants, most participants completing such a test will endorse the lowest response options for most items, leading to a positively skewed score distribution and an apparent lack of individual differences at low levels of the phenotype. If raw scores, such as the sum of items from a scale affected in this way, are used to represent the phenotype, they may provide biased tests of GxE (Molenaar & Dolan, 2014;Schwabe & van den Berg, 2014). This is because GxE estimates depend on the degrees of individual differences in a phenotype at different levels of the moderator. The use of a scale that fails to capture such differences adequately at lower levels of the phenotype may falsely indicate less variation at lower levels, when in fact this apparent observation is a function of weaker measurement at lower levels. The direction of the resulting bias in GxE depends on both skewness of the score and extent of correlation with the moderator. Positive skewness and a positive moderator-phenotype correlation is liable to produce a positive interaction parameter, while negative skewness and a positive moderator-phenotype correlation is liable to produce a negative interaction parameter.

Compounding this problem is the fact that most behaviour genetic modelling approaches require assumptions of multivariate normality[1]. With this in mind, researchers have tended to respond to observing non-normal phenotypic distributions by employing straightforward non-linear transformations intended to remove the non-normality. For

positively skewed sum scores, the log-transformation is popular (e.g Hicks, South, DiRago, Iacono, & McGue, 2009; Johnson et al., 2010) but the square root transformation is also sometimes used (e.g. Distel et al., 2011). Given that the same approach is recommended to remove GxE interactions that are artifacts of phenotypic scaling (e.g. see Falconer & MacKay, 1996 ch.17), one might conclude that this also represents a solution to the problem of dependency of GxE on scale. There are, however, at least two major reasons to doubt this. First, while there has been no systematic simulation study evaluating their effectiveness in mitigating bias due to sub-optimal scaling, Kang & Waller (2005) demonstrated that sum score transformations were only moderately successful in reducing the tendency towards spurious phenotypic interactions in the context of moderated multiple regression. Second, and more importantly: presence of GxE introduces non-normality into the phenotypic distribution because it is by definition a relative expansion or contraction of variance in the phenotype across levels of the moderator. This suggests that transforming a non-normal score to normality could 'transform away' the very interaction effect of potential interest.

As another possible solution, some authors have suggested explicitly separating out these two sources of non-normality by modelling GxE using an explicit measurement model (the scaling part) in combination with a biometric model (the GxE part). Essentially, the proposal is to model the scaling properties of items to account for differences in informativeness of phenotypic estimates across levels of the moderator. For example, if a scale has items that have difficulties that are clustered towards one end of the scale, a psychometric model with potential to recognize this can be integrated into a broader biometric model so that these parameters can be freely estimated and reflected in the estimates of the biometric parameters. The particular choice of measurement model will vary from phenotype to phenotype and be dictated by expectations about the latent trait distribution and the item response format.

For continuous indicators, Molenaar et al. (2012) demonstrated the feasibility of this approach in a GxE model in which GxE was operationalised as heteroscedastic E or C variance across levels of A. They showed that when differences in item residual variances across phenotypic level were incorporated into a measurement model and combined with a test of GxE, biasing effects of poor scaling were substantially mitigated. Similarly, Tucker-Drob et al. (2009) suggested a procedure in which a factor model with quadratic factor loadings was estimated in one stage and then, in a second stage, the same measurement model (with parameters fixed to the values estimated from the first stage) was combined with Purcell's GxE model. Quadratic factor loadings allow for the relation between the items and latent phenotype to vary across levels of the phenotype: an effect that could otherwise be mis-attributed to GxE.

However, truly continuous indicators are rare; therefore, Molenaar and Dolan (2014) and Schwabe and van den Berg (2014) proposed models for (ordered) categorical data that could be combined with a test of GxE. Again, using these models there was evidence of substantial reduction of bias in tests of GxE compared to using biometric models that did not explicitly model the scaling properties of the items used to measure the phenotype.

In spite of the potential utility of incorporating explicit measurement models for the phenotype into tests of GxE when an assumption about the underlying distribution of the genetic and environmental influences on the phenotype can be made, there have been very few studies taking this approach. One reason may be that the approach is mathematically complex and thus somewhat inaccessible for non-methodologists. There may also be a misconception that, because scores from these models will be highly correlated with sum scores, there would be essentially no benefit from using such models. It is not valid, however, to conclude that highly correlated measures will have the same properties in regression-based models, and particularly not in tests of interactions such as GxE. This is because correlations

are sensitive mainly to rank orders, which can be highly preserved even when distributional properties differ markedly. Distributional properties are particularly important in any situation involving any kind of nonlinearity such as that involved in interactions.

Misconceptions aside, there are practical limitations to the various approaches discussed above, and it is not clear what the best approach might be. For example, the Schwabe and van den Berg (2014) approach requires assumption that IRT parameters are known, the Molenaar and Dolan (2014) approach is computationally intensive, and the approaches of Molenaar et al. (2012) and Tucker-Drob et al. (2009) require continuous indicators.

Given these potential practical limitations, another possibility is to use a two-step approach to estimating GxE. In this approach, an appropriate measurement model for the phenotype is estimated and factor scores are obtained from this model, and then these factor scores are submitted to a biometric model to test GxE. The 'two steps' refer to the use of two separate models, and the approximation involved in using explicitly calculated factor scores to measure a variable conceptualized as latent. This is in contrast to the one-step approach described above in which the biometric and psychometric model are estimated together, in a single step.

Although there has been no systematic study of this approach in GxE models, simulation studies have shown that a two-step approach works well in reducing bias due to scaling in phenotypic-level interactions in moderated multiple regression and factorial ANOVA (Embreston, 1996; Kang & Waller, 2005; Morse, Johanson & Griffeth, 2012). For example, Kang and Waller (2005) showed that the tendency for spurious interactions to result from poor item scaling was substantially mitigated when IRT scores from a 2-parameter logistic model were utilised in place of sum scores. This strategy also proved more effective

than a simple non-linear transformation of the score. Therefore, it is possible that a two-step approach could provide a compromise between the greater conceptual and computational simplicity of using a sum score and the effectiveness of IRT-based latent trait estimates in accounting for the scaling properties of items.

Based on the preceding argument, I compared a two-step approach to the currently most commonly used methods for handling observed non-normal phenotypes, that is, the raw sum scores and the transformed sum scores. I compared these three approaches using a statistical simulation study complemented by a real data example.

Modelling approach

I based analyses on the GxM (gene by 'measured environment') framework initially introduced by Purcell (2002) and subsequently extended and evaluated by others (Rathouz, van Hulle, Rodgers, Waldman, & Lahey, 2008, van Hulle, Lahey, & Rathouz, 2013; Zheng & Rathouz, 2013). This framework is arguably the foremost in assessing theoretical hypotheses which predict moderation of genetic influences on a specific phenotype by a specific moderator because in addition to accommodating both gene-environment interaction and gene-environment correlation, it can also be used to evaluate a range of other forms of phenotype-moderator transactions (see Zheng & Rathouz, 2013). Uptake of the GxM modelling approach has been extensive; it has been employed to assess substantive hypotheses relating to a diversity of phenotypes including cognitive ability (Harden, Turkheimer, & Loehlin, 2007), physical health (Johnson & Krueger, 2005), health behaviours (Timberlake et al., 2006), social relationships (South, Krueger, Johnson, & Iacono, 2008), and psychopathological traits (South & Kruger, 2011). The popularity and influence of the approach is indicated by the fact that, at time of writing, the Purcell (2002) article has been cited almost 500 times.

I focussed on a form of the model that can be used to assess gene-by-measured environment interaction. The moderator (M) is modelled as:

$$M = a_M A_M + c_M C_M + e_M E_M \tag{4.1}$$

and the phenotype (P) as:

$$P = (a_C + \alpha_C M)A_M + (c_C + \gamma_C M)C_M + (e_C + \varepsilon_C M)E_M$$
$$+ (a_U + \alpha_U M)A_U + (c_U + \gamma_U M)C_U + (e_U + \varepsilon_U M)E_U, \tag{4.2}$$

where A, C and E refer to mutually uncorrelated multivariate normally distributed latent additive genetic, shared environmental and unshared environmental influences respectively, α, γ and ε are moderation parameters that capture the moderation of *A*, *C* and *E* influences by *M*, with the subscripts $_C$ and $_U$ denoting 'common' (to P and M) and 'unique' (to P).

The parameter of interest is α_U which captures the moderation of the genetic influences on the phenotype that are not shared with the moderator. When this parameter is positive, genetic influences unique to the phenotype increase with the moderator and when it is negative, they decrease with the moderator.

Simulation study

I evaluated the effect of poor scaling on estimates of α_U using Eqs. 4.1 and 4.2 as our population biometric model, simulating poor scaling of the phenotype (explained below), and

then estimating the model in Eqs. 4.1 and 4.2 using this poorly scaled phenotype. For the population biometric model, I used the following parameter magnitudes: For the moderator and phenotypic means I set $\mu_M = \mu_P = 0$; for the latent genetic and environmental influences on the moderator and phenotype I set a_U=$\sqrt{0.2}$, $a_C = \sqrt{0.3}$, $a_M = \sqrt{0.3}$; $c_U = \sqrt{0.1}$, $c_C = \sqrt{0.1}$, $c_M = \sqrt{0.2}$; e_U=$\sqrt{0.2}$, $e_C = \sqrt{0.1}$, $e_M = \sqrt{0.5}$; and for the moderation parameters I set $\alpha_C = \gamma_C = \varepsilon_C = 0$ and varied the magnitude of α_U, γ_U and ε_U across conditions. To explore how bias in α_U was affected by direction of the skewness of the observed score distribution and direction of the population interaction, I varied α_U = to be -.15, 0, and .15 across conditions. In addition, as resolvability of the α_U, γ_U, and ε_U parameters is often imperfect, I explored how the bias in α_U is affected by whether γ_U and ε_U represented interactions in the same versus the opposite direction to that of α_U. I did this by including a subset of conditions in which γ_U and ε_U were specified to have the same sign as α_U and a subset of conditions in which they were specified to have the opposite sign to α_U. In both cases the absolute magnitudes of γ_U and ε_U were specified to be .20 and .08 respectively while α_U was held constant at -.15. Together, this combination of population parameters resulted in a total of four population models, summarised in Table 4.2. In each replication, I generated data for 500 MZ and 500 DZ twins according to these models.

I selected parameter magnitudes representing realistic values from previous empirical studies. Because results could be expected to be broadly symmetrical for positive and negative skews and negative and positive interaction parameters, I did not implement a fully crossed simulation design, but focussed on models that were realistic and which covered key combinations of variables.

Observed data generation

I generated item level data for twin 1 and twin 2 separately using a graded response model (GRM; Samejima, 1969) as the basis for linking the latent trait values for the phenotype (P) to observed item responses. An identical GRM model was used for twin 1 and twin 2. This allowed the same model to apply to all individuals in the sample while also having the advantage of allowing any complications due to clustering within twin pairs to be circumvented. These latent trait values were determined according to the GxM population models described in the previous section. I simulated these data using the catIrt package in R statistical software (Nydeck, 2014; R Core Team, 2014). In the GRM, the items are essentially considered in dichotomous steps, each characterised by a 2-parameter logistic model but with discriminations constrained equal within items. Specifically, probability of a respondent *i* with level of the latent trait θ_i having a response x_{ij} that falls at or above a given category ($k = 1 \ldots m_j$) is specified as:

$$P^*_{ijk} = P\left(x_{ij} \geq k \middle| \theta_i, a_j, \beta_{jk}\right) = \frac{1}{1 + \exp[-\alpha_j(\theta_i - \beta_{jk})]} \quad (4.3)$$

where α_j is the discrimination parameter of item j and β_{jk} is the category difficulty parameter of category *k* in item *j*.

I generated data for 20 items with α_j and β_{jk} parameters provided in Table 4.1.

Table 4.1: Parameter values for IRT model used in simulation

Item	$\boldsymbol{a}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$	$\boldsymbol{\beta_4}$
1	1.94	-0.27	0.84	2.23	2.74
2	1.93	-0.21	1.46	2.01	2.73
3	1.96	-0.11	1.50	2.38	2.82
4	2.13	-0.36	1.29	2.07	2.65
5	1.09	0.34	1.16	2.07	2.73
6	1.13	-0.15	1.34	2.00	2.78
7	0.87	0.34	0.99	2.34	2.64
8	0.99	0.23	0.68	2.33	2.62
9	1.63	0.43	0.98	2.22	2.83
10	1.01	0.04	1.22	2.39	2.73
11	1.75	0.10	0.93	2.27	2.63
12	0.80	0.01	0.67	2.20	2.75
13	0.67	0.37	1.49	2.42	2.67
14	1.91	0.13	0.89	2.29	2.92
15	1.06	0	1.29	2.09	2.96
16	0.55	0.50	0.76	2.32	2.81
17	1.88	-0.24	1.02	2.07	2.74
18	2.44	-0.40	0.80	2.09	2.86
19	0.90	-0.11	1.27	2.27	2.73
20	1.15	-0.24	0.65	2.17	2.73

Note. $\boldsymbol{a}$ is an item discrimination parameter, $\boldsymbol{\beta_1}$ -$\boldsymbol{\beta_4}$ are threshold parameters.

The β_{jk} parameters were chosen to yield positively skewed item and sum score distributions that mimicked those commonly found in empirical research (e.g. Kang & Waller, 2005). To do this, I selected β_{jk} for successive response categories so that a

disproportionate number of responses would fall into the first and second response categories. I also specified the β_{jk} parameters for a given category to show variability across the 20 items within our simulated test which is more realistic than setting them all equal. Discrimination parameters, α_j, were selected by randomly sampling from a uniform distribution with min = 0.5 and max = 2.5.

True score

As a control condition, I generated scores for the phenotype according to Eqs. 4.1 and 4.2 without introducing any scaling issues. These scores can therefore be considered 'true' phenotypic scores. I considered these true phenotypic scores in order to provide a baseline against which I could compare the results. This is necessary because even in the absence of any scaling problems, it is likely that the GxM model will not perfectly recover all moderation parameters and because moderation parameters may be difficult to resolve from one another. For example, moderation of shared environmental influence may be to some extent mis-attributed to moderation of genetic influences.

Sum score

I created sum scores for the phenotype summing the scores from the 20 items generated as described above by Eqs 4.1, 4.2, and 4.3. An example of a resulting sum score distribution is shown in Figure 4.1. It illustrates that the choice of GRM parameters in Table 4.2 yielded a sum score distribution exhibiting moderate positive skewness, similar to that observed in many measures of psychopathological traits. Skewness also depended on the direction of the interaction in the population model, with positive interactions making score distributions more positively skewed and negative interactions making score distributions more negatively skewed. However, these effects were relatively minor in comparison to the effect of scaling on the phenotypic distribution.

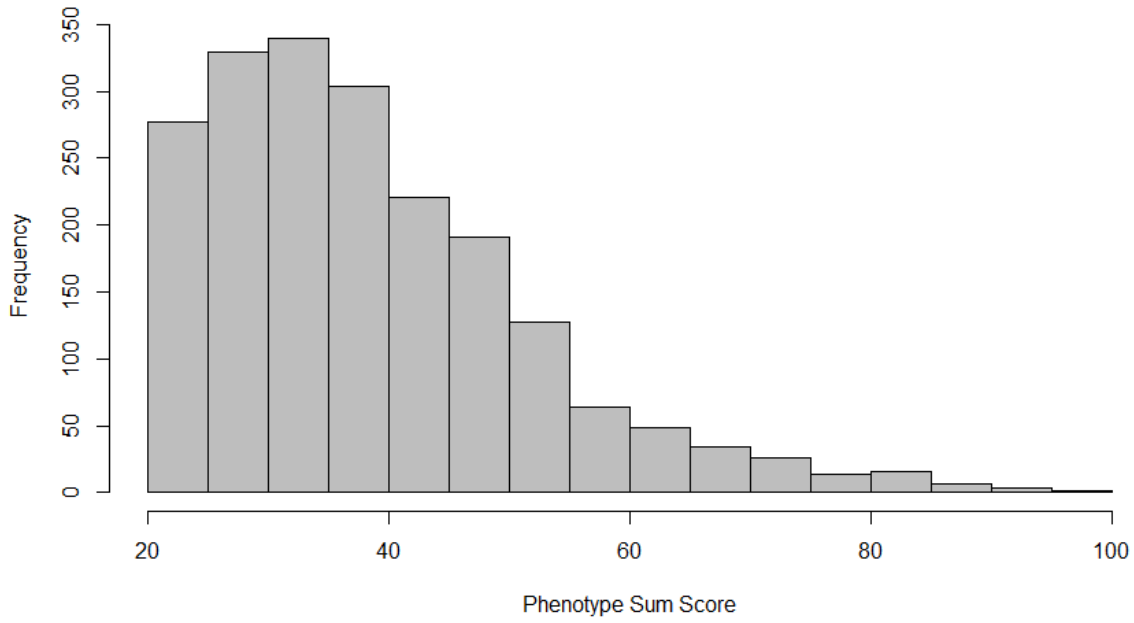


Figure 4.1

Histogram showing the distribution of the sum score derived from generating item level data according to Eq. 4.3 with parameters in Table 4.1.

Transformed sum score

I created transformed sum scores for the phenotypes using a $\log_{10}$ transformation. This, the natural log transformation and other similar kinds of transformations of the phenotype are commonly used in GxE models when the phenotype has a positively skewed distribution (e.g. Button et al., 2010; Hicks, Dirago, Iacono, & McGue, 2009; Hicks et al., 2009; Johnson et al., 2010; Silvetoinen et al., 2009; Tuvblad, Grann, & Lichtenstein, 2006). Transforming the sum scores gave rise to approximately normal distributions (see Figure 4.2).

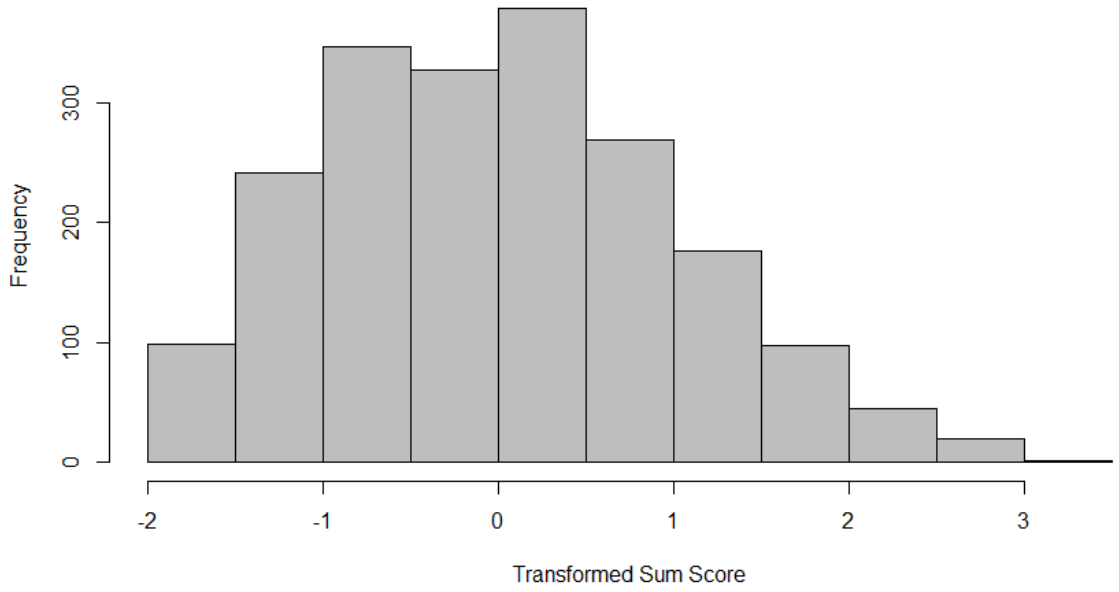


Figure 4.2

Histogram showing the distribution of the transformed sum score derived from generating item level data according to Eq. 4.3 with parameters in Table 4.1 and then applying a $\log_{10}$ transformation.

IRT scores

I obtained factor scores by fitting an IRT model to the item data and using the resulting item parameters to estimate IRT-based individual phenotype scores, usually referred to as 'factor scores' (Chalmers, 2012). To estimate item parameters, I fit a graded response model to the data. As I originally generated the data according to a graded response model, I knew this was the appropriate measurement model, however, in real applications this choice should be based on considerations of the response format of items and the likely form of relations between item responses and the latent phenotype. I then computed IRT-based estimates of the phenotypic level for each individual in the sample by combining information from their patterns of item scores with the estimated item parameters from fitting the graded

response model, specifically, using Expected a Posteriori (EAP) scoring (Embretson & Reise, 2000). EAP scoring is a Bayesian approach based on finding the mean of a posterior distribution representing the likelihood of phenotypic scores given a response pattern. The posterior distribution is computed by multiplying the prior distribution (likelihoods of phenotypic levels occurring in the population) by the likelihood of the observed response pattern given the phenotypic level (Embretson & Reise, 2000). This method was selected among available factor score estimation approaches because it is easy to implement and available in most IRT software packages. In context of the models used here in which the trait of interest is uni-dimensional and the sample size large, I anticipated that other commonly used scoring methods such as maximum a posteriori (MAP) scoring or maximum likelihood estimates (ML) would perform similarly to EAP. Unlike using sum scores as a proxy for the phenotype, this method takes into account the scaling properties of the items. For example, in an IRT model in which items differ in discrimination, each item's contribution to the sum score will depend on its discrimination. Estimating factor scores in this way gave phenotypic scores with an approximately normal distribution (see Figure 4.3).

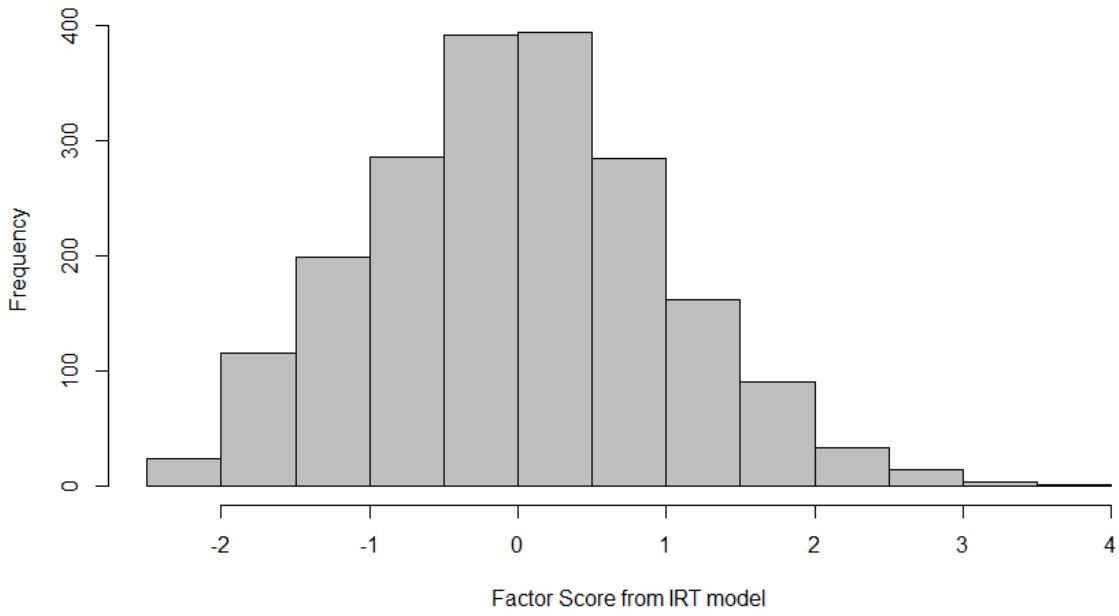


Figure 4.3

Histogram showing the approximate distribution of factor scores derived from generating item level data according to Eq. 3 with parameters in Table 4.1, fitting a graded response model, and then obtaining factor scores based on this model.

Summary of simulation conditions

The combination of GxE interaction parameters (α_U, = -.15 vs 0 vs .15), other interaction parameters (γ_U = .20 and ε_U = .08 vs γ_U = -.20 and ε_U = -.08), and score type (true, sum, transformed, IRT) resulted in 16 simulation conditions. These are outlined in Table 4.2. I generated 100 datasets for each condition to give 100 replications per condition.

Model fitting

To the 100 simulated datasets for each simulation condition (see Table 4.2), I fit the GxM model described in Eqs. 1-2. I fit the models in Mx (Neale, Boker, Xie, & Maes, 2006) using maximum likelihood estimation, making use of the script accompanying Purcell (2002)

which the author has made available on his website (http://pngu.mgh.harvard.edu/~purcell/gxe/). All latent A,C and E variances and covariances were freely estimated, α_C, $\boldsymbol{\gamma}_C$, and $\boldsymbol{\varepsilon}_C$ were fixed to zero, and α_U, $\boldsymbol{\gamma}_U$ and $\boldsymbol{\varepsilon}_U$ were freely estimated. In other words, the model I fit to each dataset was consistent with the true model. The main parameter of interest was α_U, which captures the moderation of the additive genetic variance unique to the phenotype by M. However, because this parameter may not be completely resolvable from $\boldsymbol{\gamma}_U$ and $\boldsymbol{\varepsilon}_U$, I recorded the mean and SD estimates across the 100 replications for all three moderation parameters across each condition. Parameter bias was the difference between the population magnitude and the mean estimated value across the 100 replications within a condition. In addition, I conducted a likelihood ratio test (comparing a model in which α_U was freely estimated to one in which it was constrained to zero) for each replication to evaluate the statistical significance of the α_U, parameter. Based on these, I computed false positive and false negative rates across the 100 replications. False negative rate was defined as the proportion of replications in which α_U, was non-significant in the presence of a non-zero population parameter. False positive rate was defined as the proportion of replications in which: a) α_U was significant in the presence of a null population parameter or b) α_U was statistically significant but its value was in the opposite direction to its population value (e.g. negative sample value with a positive population value).

Simulation Study Results

Simulation study results are provided in Table 4.2. For the 'true scores', the α_U parameters were generally recovered well. Power to detect moderation was high and greatest when it was in the same direction as the main effects and the $\boldsymbol{\gamma}_U$ and $\boldsymbol{\varepsilon}_U$ parameters.

Sum scores conditions

In all conditions in which a poorly scaled sum score was used as the phenotype ('sum score' rows of Table 4.2), there was positive bias in the α_U parameter, with $\boldsymbol{\gamma}_U$ and $\boldsymbol{\varepsilon}_U$ also tending to be affected in the same way. The positive biases occurred because the IRT parameters used to generate the data produced positively skewed sum scores. Had item parameters been selected to produce a negatively skewed sum scores, negative biases would have occurred. Positive α_U bias was largest in conditions in which the true moderation parameter was in the opposite direction from the direction of skew (i.e. a negative or null population moderation parameter with a positively skewed score) and the other moderation parameters. Here the biasing effects of scaling and imperfect resolvability of the α_U and $\boldsymbol{\gamma}_U$ parameters seemed to show effects which combined to give a larger overall positive bias. The false positive rate using sum scores was also high. This suggests that significant moderation detected using poorly scaled sum scores cannot not be trusted.

Table 4.2: Performance of sum score, transformed score and IRT score latent trait proxies under different population biometric models

Score type	Population GxM values						Convergence failures (%)	Average α_U (SD)	α_U Bias	α_U true positive rate	α_U false positive rate[a]	Average γ_U (SD)	γ_U bias	Average ε_U (SD)	ε_U bias
	a_c	c_c	e_c	α_U	γ_U	ε_U									
True	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	.15	.20	.08	0	0.15 (0.04)	0.00	98%	0%	0.18 (0.05)	-0.01	0.08 (0.02)	0.00
True	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	-.15	.20	.08	0	-0.12 (0.05)	+0.03	75%	0%	0.19 (0.04)	-0.01	0.07 (0.01)	-0.01
True	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	0	.20	.08	0	0.00 (0.03)	0.00	N/A	0%	0.19 (0.03)	-0.01	0.08 (0.02)	0.00
True	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	-.15	-.20	-.08	0	-0.15 (0.05)	0.00	96%	0%	0.16 (0.08)	-0.04	-0.08 (0.02)	0.00
Sum	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	.15	.20	.08	0	0.22 (0.05)	+0.07	94%	0%	0.17 (0.07)	-0.03	0.15 (0.02)	+0.07
Sum	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	-.15	.20	.08	0	0.03 (0.08)	+0.18	9%	1%	0.21 (0.05)	+0.01	0.15 (0.02)	+0.07
Sum	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	0	.20	.08	0	0.14 (0.07)	+0.14	N/A	54%	0.18 (0.08)	+0.02	0.15 (0.02)	+0.08
Sum	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	-.15	-.20	-.08	0	-0.06 (0.05)	+0.09	15%	0%	0.00 (0.10)	+0.20	0.05 (0.02)	+0.13
Transformed	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	.15	.20	.08	0	0.16 (0.03)	+0.01	73%	0%	0.15 (0.04)	-0.05	0.06 (0.01)	-0.02
Transformed	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	-.15	.20	.08	0	-0.02 (0.05)	+0.13	4%	0%	0.11 (0.05)	-0.09	0.06 (0.02)	-0.02
Transformed	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	0	.20	.08	0	0.08 (0.04)	+0.08	N/A	23%	0.12 (0.04)	-0.08	0.05 (0.02)	-0.03
Transformed	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	-.15	-.20	-.08	1	-0.11 (0.03)	+0.04	68%	0%	-0.08 (0.07)	+0.12	-0.05 (0.02)	+0.03
IRT	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	.15	.20	.08	0	0.16 (0.04)	+0.01	80%	0%	0.15 (0.05)	-0.05	0.05 (0.01)	-0.03
IRT	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	-.15	.20	.08	0	-0.06 (0.05)	+0.09	13%	0%	0.16 (0.04)	-0.04	0.06 (0.02)	-0.02
IRT	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	0	.20	.08	0	0.06 (0.05)	+0.02	N/A	16%	0.14 (0.05)	-0.06	0.06 (0.02)	-0.02
IRT	$\sqrt{.3}$	$\sqrt{.1}$	$\sqrt{.1}$	-.15	-.20	-.08	0	-0.13 (0.03)	+0.02	79%	0%	-0.01 (0.13)	+0.19	-0.05 (0.02)	+0.03

[a]False positive rate defined as significant effect in opposite direction to population parameter or significant effect in any direction when population parameter is zero.

Transformed sum scores conditions

Overall, there was positive bias in α_U using transformed sum scores, especially in the condition in which the true moderation parameter was in the opposite direction to both the direction of skew and the other moderation parameters. The effect of transforming the sum score to normality was to pull the α_U parameter in the negative direction. This represented a reduction in bias relative to using the untransformed sum score. Effectiveness of the transformation for reducing bias varied across conditions: it almost eliminated the mild bias in the condition in which all moderation parameters were in the same direction as the scaling effects but moderate to substantial positive bias remained in the other conditions and GxE was under-estimated. Under-estimation of GxE effects meant that power to detect GxE was substantially reduced, particularly in the condition in which the GxE was in the opposite direction to the main and other moderation effects. Here, the true positive rate dropped from 75% for the true scores to only 4% for the transformed sum scores. However, transforming the sum scores to normality had the benefit of producing a marked reduction in false positive rate. When the population parameter was zero the false positive rate was only 23% when using a transformed sum score, compared with 54% when using a raw sum score.

IRT scores conditions

Overall, using factor scores from an appropriate IRT model as the phenotypic proxy (rows labelled 'IRT score' in Table 4.2) gave less biased α_U parameter estimates than either raw or transformed sum scores., however, some positive bias remained in all cases. This bias was most pronounced in the condition in which the α_U parameter was in the opposite direction to the scaling problems and the other moderation parameters (+.09) but minimal (+.01 to +.02) in the other conditions. The power to detect GxE was lower when using an IRT score in the corresponding true score conditions but higher than when using a transformed sum score. The false positive rate also compared favourably to that obtained using a

transformed sum score (16% compared with 23% in when using transformed sum scores) but unfortunately, remained above nominal levels (i.e. 5%).

Real Data Example

Participants

To provide a real data example, I used data from the Minnesota Twin Registry (MTR), a comprehensive description of which can be found in Krueger and Johnson (2004). The full MTR includes data from twin pairs born in Minnesota in one of three year ranges. It includes 4307 twin pairs born between 1936 and 1955, 901 twin pairs born between 1904 and 1943, and 391 male twin pairs born between 1961 and 1964. Eligible participants were identified from birth records, located, and invited to participate via mail. Additional incentives and invitations to participate were offered to those who did not initially respond. Zygosity determination was by self-reported similarity in eye colour, hair colour, overall appearance, and the difficulties others had in distinguishing two members of a pair. Analysis of a sub-sample of 74 twin pairs who underwent zygosity determination by serological analysis suggested that the self-report method had an estimated accuracy of 96%.

Different subsets of the total MTR received different sets of measures. Data used in the current study were from 528 monozygotic twin pairs and 411 dizygotic twin pairs comprising 614 males and 1264 females who had completed measures of both personality and leisure time interests. The mean age of the sample was 37.11 (SD = 7.8).

Measures

Moderator

As the moderator variable, I used a composite of items from the Minnesota Leisure Time Interest Test (MLTIT; Lykken et al., 1990). The scale asks participants to rate the

extent to which they would be interested in pursuing a given activity assuming no time, health, or financial constraints. Participants rated their interest on a 5-point scale from 1 = 'No interest at all' to 5 = 'I would certainly do this'. In total, 120 activities were rated, but I selected 6 items to form an 'Intellectual Interests' scale. Selected items refer to the following activities: reading current non-fiction, taking a college course, reading literary classics, visiting galleries/museums/exhibitions, reading books/magazines or watching TV programs on science, and reading history/philosophy/biography. I checked that these items formed a reasonable uni-dimensional scale by fitting a single factor confirmatory factor model to the data from twin 1 of each twin pair. I used the Weighted Least Squares Means and Variances (WLSMV) in estimator in *Mplus 7.0* (Muthén & Muthén, 2010) to account for the categorical item response format. Fit statistics and parameter estimates suggested that, according to conventional criteria, it would be reasonable to combine the items into a single scale: the 6 items all showed standardised loadings of .50 or greater and yielded a good-fitting single factor model (RMSEA = .05, CFI = .99, TLI = .99, WRMR = 0.56).. I therefore used the unweighted sum score of these six items as our moderator variable. Cronbach's alpha of the scale was .63.

Phenotype

As the phenotypes I used personality scales from the 300-item Multidimensional Personality Questionnaire (MPQ; Tellegen & Waller, 2008). Participants were administered a version of the MPQ using a 2-point response scale. Items are phrased as statements to which participants answer 'True' or 'False' depending on whether they believe the statement describes their attitudes, opinions, interests or other characteristics.

I selected two scales that yielded oppositely skewed scores. First, I used the negatively skewed 'Well-being' scale comprising 18 items. High scores on this scale are

presumed to be indicative of a cheerful and happy disposition, feeling good about oneself, being optimistic, and enjoying an interesting and exciting life. Second, I used the positively skewed 'Aggression' scale comprising 18 items. High scores on this scale are presumed to be indicative of physical aggression, enjoyment of scenes of violence or upsetting or frightening others, victimisation of others for personal advantage, and vindictive and retaliatory tendencies.

I varied how each phenotype was operationalised across conditions to mirror our simulation conditions. First, I used the raw sum score from each scale. Second, I used a transformation of the sum score that yielded an approximately normal distribution. Third, I used an IRT score for each scale. For this, I used a 2-parameter logistic model with a procedure otherwise identical to that described in the simulation study to estimate factor scores.

Model fitting

Model fitting broadly followed the procedure outlined in the simulation. However, because I was working with real data, I did not know the true model and, therefore, relied on model fit comparisons to guide model selection. I first assessed whether it was possible to constrain moderation of the influences common to moderator and phenotype to zero without significant decrease in fit. I then attended to moderation of the influences unique to the phenotype. I present the parameter estimates from best-fitting model(s). In all cases, all latent A, C, and E variances and covariances were freely estimated.

Real Data Example Results

Descriptive Statistics

Descriptive statistics for the moderator and phenotypes are provided in Table 4.3. For the phenotypes, descriptive statistics are provided for sum scores, transformed sum scores and IRT scores. The Well-being sum score showed negative skew which was reduced considerably by a normalising transformation, specifically, a squaring of scores. The IRT factor scores for this phenotype showed a level of non-normality similar to the transformed sum score but slightly more negative. The correlation between Well-being and Intellectual interests was around r=.18 and practically unaffected by which phenotypic proxy was used. The correlations between the three kinds of scores derived from the Well-being items were all >.97.

The Aggression sum score showed positive skewness. Given the magnitude of positive skewness, a natural log transformation was used and this produced scores with a near-normal distribution. The IRT factor scores for this phenotype also substantially reduced non-normality but these scores were more positively skewed than the transformed sum scores. The correlation between Aggression and Intellectual interests was around r=-.12 and practically identical across the three different kinds of phenotypic proxy. The correlations between the three kinds of scores derived from the Aggression items were also all >.97.

Table 4.3: Descriptive statistics for Well-being, Aggression and Intellectual Interests phenotypes

Phenotypic proxy	Mean (SD)	Skew	Kurtosis	Correlation with moderator
Intellectual Interests sum score	13.32 (3.75)	0.13	-0.27	N/A
Well-being sum score	11.15 (2.21)	-1.06	0.71	.18
Well-being sum score transformed	0 (1)	-0.36	-0.90	.19
Well-being IRT score	0 (0.89)	-0.42	-0.32	.18
Aggression sum score	3.66 (3.21)	1.12	1.09	-.12
Aggression sum score transformed	0 (1)	0.23	-0.79	-.12
Aggression IRT score	-0.04 (0.86)	0.46	-0.40	-.13

Well-being

Model fits for the Well-being scale GxM Models are provided in Table 4.4. Parameter estimates for the best fitting model are provided in Table 4.5.

Table 4.4: GxM model fits for Well-being phenotype

Model (freely estimated parameters)	-2LL	Df	BIC	AIC	saBIC	DIC
			Sum score			
$a_C, c_C, e_C, \alpha_C, \gamma_C, \varepsilon_C, \alpha_U, \gamma_U, \varepsilon_U$	10204.50	3727	-7653.07	2750.50	-1734.73	-4228.18
$a_C, c_C, e_C, \alpha_U, \gamma_U, \varepsilon_U$	10204.78	3730	-7663.18	2744.80	-1740.08	-4235.54
a_C, c_C, e_C, α_U	10206.10	3732	-7669.38	2742.09	-1743.11	-4239.91
a_C, c_C, e_C	10222.75	3733	-7664.47	2756.75	-1736.61	-4234.08
			Transformed sum score			
$a_C, c_C, e_C, \alpha_C, \gamma_C, \varepsilon_C, \alpha_U, \gamma_U, \varepsilon_U$	10214.25	3727	-7648.19	2760.25	-1729.85	-4223.30
$a_C, c_C, e_C, \alpha_U, \gamma_U, \varepsilon_U$	10214.92	3730	-7658.12	2754.92	-1735.02	-4230.48
a_C, c_C, e_C, α_U	10215.09	3732	-7664.88	2751.09	-1738.60	-4235.40
a_C, c_C, e_C	10219.96	3733	-7665.87	2753.96	-1738.00	-4235.47
			IRT score			
$a_C, c_C, e_C, \alpha_C, \gamma_C, \varepsilon_C, \alpha_U, \gamma_U, \varepsilon_U$	9806.21	3739	-7893.28	2328.21	-1955.88	-4457.37
$a_C, c_C, e_C, \alpha_U, \gamma_U, \varepsilon_U$	9806.89	3742	-7903.21	2322.89	-1961.05	-4464.54
a_C, c_C, e_C, α_U	9807.08	3744	-7909.96	2319.09	-1964.62	-4469.45
a_C, c_C, e_C	9810.82	3745	-7911.51	2320.82	-1964.59	-4470.08

For the Well-being scale, the correlations between the three types of score were all >.97. The negative skew of the raw sum score was markedly reduced in both the transformed sum score and the IRT factor score; however, there was effectively no difference in their correlation with the moderator. This illustrates the important point that highly correlated scores or scores with effectively identical correlations with the moderator will not necessary be equivalent with respect to the distributional properties that GxE tests are sensitive to.

In the GxE models for this phenotype, it was possible to constrain moderation of the common influences to zero without significant decrease in fit irrespective of whether a sum score, transformed sum score, or IRT score represented the phenotype. Therefore, this became the baseline model for all further model comparisons.

Using sum scores, model comparisons supported moderation of the genetic influences unique to the phenotype fairly unequivocally. Constraining this parameter to zero produced significant decrease in fit irrespective of whether moderation of the unique C and E influences on the phenotype was freely estimated or fixed to zero. Model fit comparisons suggested the latter model provided the best overall representation of the data. Thus, results suggested that the genetic influences unique to Well-being were smaller at higher levels of intellectual interests.

Using transformed sum scores, model fit comparisons suggested some moderation of unique influences for which moderation of the A influences unique to the phenotype best accounted. However, this result was not completely unequivocal: it was possible to constrain moderation of the A influences unique to the phenotype to zero without significant decrease in fit when moderation of the C and E influences were freely estimated but not when they were both fixed to zero. This further illustrates the lack of resolvability of α_U and γ_U effects noted in the simulation study. Here results suggested that the genetic influences unique to Well-being may be higher at higher levels of intellectual interests.

When using IRT scores, results were highly similar to those for the transformed sum score in terms of fit differences and parameter magnitudes (α_U was 0.04 when freely estimated but the other moderation parameters were fixed to zero). However, the difference in fit between the model in which moderation of all the unique A, C and E influences on the phenotype was fixed to zero and the model in which moderation of the unique A influences was freely estimated happened to fall just short of statistical significance. Therefore, there was technically no statistical evidence for GxE when using the IRT factor score, suggesting that the genetic influences unique to Well-being did not depend on level of intellectual Interests.

To summarise results from the Well-being scale, based on a naïve interpretation, all favoured different conclusions regarding the presence of GxE: GxE was in evidence using a sum score, was somewhat in evidence using a transformed sum score, and was not in evidence using an IRT score. While the results in the latter two conditions were in actuality very similar, the fact that the statistical evidence lay on opposite sides of a statistical significance threshold and a naïve interpretation could lead to very different substantive conclusions in practice. Only the sum score condition appeared to show unambiguous support for GxE. This is consistent with the simulation conditions in which the presence of non-normality resulted in detection of GxE, irrespective of whether this non-normality was a result of moderation or poor scaling. The moderation observed using the sum score was in the direction expected for a negatively skewed sum score even when there was no true moderation. Thus, there would be reason to question the validity of the evidence for GxE observed in this real data example.

Table 4.5: Parameter estimates from best-fitting models for Well-being phenotype

	Phenotype				**GxM Parameter Estimates**	
Phenotypic Proxy	α_C	α_U	γ_C	γ_U	ε_C	ε_U
Sum score	0 (fixed)	-.11	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)
Transformed sum score	0 (fixed)	-.06	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)
IRT factor score	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)

Aggression

GxM model fits for the aggression scale are provided in Table 4.6 and parameter estimates from the best fitting model are provided in Table 4.7. For this phenotype, the correlations between the three kinds of scores were also all >.97. The raw sum score showed

substantial positive skew but both the transformed sum score and the IRT score had reasonably symmetrical distributions. Again, however, the correlations of the phenotype with the moderator were practically identical, irrespective of which score type was used and in spite of the marked differences in their distributions.

Table 4.6: GxM model fits for Aggression phenotype

Model (freely estimated parameters)	-2LL	df	BIC	AIC	saBIC	DIC
Sum score						
$a_C, c_C, e_C, \alpha_C, \gamma_C, \varepsilon_C, \alpha_U, \gamma_U, \varepsilon_U$	10218.91	3732	-7662.97	2754.91	-1736.69	-4233.49
$a_C, c_C, e_C, \alpha_U, \gamma_U, \varepsilon_U$	10222.38	3735	-7671.51	2752.38	-1740.46	-4239.27
$a_C, c_C, e_C, \varepsilon_U$	10224.28	3737	-7677.40	2750.28	-1743.18	-4243.33
a_C, c_C, e_C	10240.40	3738	-7672.76	2764.40	-1736.96	-4237.77
Transformed sum score						
$a_C, c_C, e_C, \alpha_C, \gamma_C, \varepsilon_C, \alpha_U, \gamma_U, \varepsilon_U$	10228.85	3732	-7658.00	2764.85	-1731.72	-4228.52
$a_C, c_C, e_C, \alpha_U, \gamma_U, \varepsilon_U$	10232.34	3735	-7666.52	2762.34	-1735.48	-4234.29
$a_C, c_C, e_C, \varepsilon_U$	10234.73	3737	-7672.17	2760.73	-1737.96	-4238.10
a_C, c_C, e_C	10238.00	3738	-7673.96	2762.00	-1738.16	-4238.97
IRT score						
$a_C, c_C, e_C, \alpha_C, \gamma_C, \varepsilon_C, \alpha_U, \gamma_U, \varepsilon_U$	9676.16	3739	-7958.30	2198.16	-2020.91	-4522.39
$a_C, c_C, e_C, \alpha_U, \gamma_U, \varepsilon_U$	9679.97	3742	-7966.67	2195.97	-2024.51	-4528.00
$a_C, c_C, e_C, \varepsilon_U$	9682.21	3744	-7972.39	2194.21	-2027.06	-4531.88
a_C, c_C, e_C	9687.08	3745	-7973.38	2197.08	-2026.46	-4531.95

In all conditions, it was possible to constrain moderation of the influences common to moderator and phenotype to zero without significant drop in fit. From here, the best-fitting model in using a sum score was one in which there was moderation of the unshared environmental influences on the phenotype. When this parameter was freely estimated, constraining moderation of neither shared environmental influences nor genetic influences on the phenotype resulted in statistically significant decrease in fit. Thus, using a sum score,

there was evidence that only the unshared environmental influences unique to Aggression decreased with increasing intellectual interests. The direction of this moderation was in the opposite direction to the direction of the skew of the sum score. Given that the phenotype and moderator were negatively correlated, the moderation was in the direction consistent with the skew of the sum score.

Using transformed sum scores, after constraining moderation of the influences common to moderator and phenotype to zero, the best-fitting model involved no moderation of the influences unique the phenotype. These could all be individually constrained to zero without significant decrease in fit, irrespective of whether moderation parameters for the other unique influences were also constrained or freely estimated. Thus, there was no evidence that the genetic or environmental influences on Aggression depended on level of Intellectual Interests.

Using IRT scores, after constraining moderation of the influences common to the moderator and phenotype to zero, there was some very weak support for moderation of the unshared environmental influences unique to the phenotype. Specifically, fixing moderation of unshared environmental influences unique to the phenotype to zero resulted in significant decrease in fit when all other moderation parameters were fixed to zero. Further, the decrease in fit on constraining this parameter to zero was not statistically significant when moderation of the shared environmental and genetic influences unique to the phenotype was freely estimated. In addition, the best-fitting model according to BIC included no moderation, albeit by a small margin compared with one in which the moderation of the unshared environmental influences unique to the phenotype was freely estimated (ΔBIC = 0.99). Therefore, on balance the IRT factor score condition showed only very weak evidence for moderation intermediate between the results for the sum score (which showed evidence for

moderation) and the transformed sum score (which showed no evidence for moderation). Again, the direction of moderation suggested smaller unshared environmental influences.

Table 4.7: Parameter estimates from best-fitting models for Aggression phenotype

	Phenotype		GxM Parameter Estimates			
Phenotypic Proxy	α_C	α_U	γ_C	γ_U	ε_C	ε_U
Sum score	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	-0.07
Transformed sum score	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)
IRT factor score	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	-0.03

Discussion

It is well known that using poorly scaled sum scores as phenotypic proxies in GxE tests can seriously bias tests of GxE. For example, using sets of items where the difficulty or location parameters are clustered near the low end of the phenotypic continuum can lead to positively skewed sum scores and, in turn, positively biased tests of GxE (e.g. Molenaar & Dolan, 2014; Schwabe & van den Berg, 2014). In a simulation study, I assessed the extent to which this bias was mitigated by transforming non-normal sum scores to normality. I compared this to estimating phenotypic scores from an IRT model: a method that explicitly takes account of the scaling properties of items. The results suggested that using IRT methods to provide formal models for the phenotype is worth the effort in providing more accurate detection and quantification of GxE effects.

Based on these analyses, I can extend the arguments set out in the introduction in the following ways. First, I confirmed that biases in estimates of GxE can be introduced by poor phenotypic scaling that result in sum scores that do not accurately reflect the underlying

distributions of the phenotypes they are supposed to represent. The nature of this bias is predictable: sum scores that are negatively skewed relative to their underlying phenotypic distribution will tend to produce negatively biased moderation parameters and sum scores that are positively skewed relative to their underlying phenotypic distribution will tend to produce positively biased moderation parameters. When there is no true moderation effect, this will often lead to unacceptably high false positive rates.

These effects occur because non-normality due to poor scaling is not statistically distinguishable from non-normality due to presence of interaction. Where there is non-normality, the model will attribute this to interaction; however, only when the observed phenotypic distribution reflects its population distribution will this estimate provide accurate quantification of GxE. Measuring the phenotype and capturing its population distribution as accurately as possible is, therefore, important in ensuring accurate assessment of GxE. When the raw score from an inventory fails to do this, there may be options for recovering this distribution via *post-hoc* manipulations of its measurement scale.

Results showed, in particular, that transforming a score or using an IRT score in place of a non-normal sum score can be used to reduce bias. I studied the case in which the *latent* genetic and environmental influences on the phenotype, absent the influence of the moderator could be assumed normally distributed in the population. This is a reasonable assumption in cases where there are large numbers of small, independent effects on the phenotype. Here, a normal distribution of the joint effects of numerous relatively fungible etiological contributors is predicted based on the central limit theorem. Under these conditions, using either a simple transformation or IRT scores reduced bias in GxE because they led to score distributions that better approximated the population distribution of the phenotype.

In a case where there is no true moderation effect, using a phenotypic proxy that better reflects its population distribution than a sum score reduces false positive rates substantially. When the direction of the moderation is consistent with the direction of skew, either transforming to normality or using an IRT score will give close to unbiased parameter estimates and result in good power to detect the effect. However, in cases where moderation and skew are in opposite directions, these methods will underestimate the effects and reduce power to detect GxE relative to situations in which the phenotypes are not subject to scaling problems.

I also provided a real data example from the Minnesota Twin Registry using two phenotypes with non-normal sum scores. Analysing the Well-Being phenotype using (negatively skewed) sum scores yielded statistically and practically significant GxE whereas using IRT scores suggested no significant GxE. The transformed sum scores yielded evidence intermediate between these two outcomes. The direction of the GxE using sum scores was consistent with the direction of the skewness of the sum score. This suggests that the observed effect could be due to item scaling. Moreover, based on these results, researchers using sum scores rather than IRT scores could easily have been led to opposite substantive conclusions despite the very high correlations between the raw and IRT scores.

The Aggression phenotype did not yield evidence of GxE irrespective of whether (positively skewed) sum scores, transformed sum scores, or IRT scores were used. This showed that non-normal trait distributions will not automatically result in the appearance of GxE and that altering phenotypic distributions will not necessarily affect the GxE parameter. However, there was evidence for dependence of another moderation parameter on scaling: using a sum score and an IRT score, there was evidence for negative moderation of the unshared environmental influences unique to the phenotype (captured by the $\boldsymbol{\varepsilon}_U$ parameter) but there was no such evidence using a transformed sum score. Taking into account the fact

that the phenotype and moderator were negatively correlated, the $\boldsymbol{\varepsilon}_U$ parameter was proportional to and in the direction consistent with the skew of the phenotypic proxy. That is, the parameter was most negative when the phenotypic proxy was strongly skewed (sum score), less negative when the phenotypic proxy was moderately positively skewed (IRT score) and effectively zero when the phenotypic proxy was only slightly positively skewed (transformed sum score). Thus, although I focussed on the α_U parameter because it is most often used to operationalise theoretical hypotheses, this example highlights the fact that the effects of scaling on GxE models are not confined to that one parameter.

These results reinforce the message that poorly scaled sum scores should be avoided in tests of GxE. However, they also imply that the commonly used strategy of transforming non-normal sum scores to normality will in many cases fail to address the biasing effects of poor scaling on GxE tests fully, particularly when the true moderation parameter is in the opposite direction to both scale skew and moderation of the unshared and shared environmental influences on the phenotype. IRT scores were also subject to this limitation but overall performed better than transforming the sum score to normality in terms of parameter bias, false positive rates and true positive rates.

Practically speaking, sum scores are not desirable as phenotypic proxies because, in addition to producing high false positive rates, they can yield results that suggest moderation in the opposite direction to the true moderation effect. The strategies of transforming the sum score to normality or using an IRT score do not suffer these limitations; however, both result in tests that lack statistical power when the moderation is in the opposite direction to skew. Given this, transformed sum scores and IRT scores provide conservative tests of GxE when it is present. However, they also fail to control the type 1 error rate completely when GxE is not present; therefore, caution is still due when interpreting significant GxE tests obtained using these scores.

Demonstrating that sum scores are highly correlated with transformed sum scores or IRT scores for the same phenotype is thus not sufficient justification for using them in place of these better-performing methods. Because correlation coefficients are relatively unaffected by rank-preserving transformations, sum and functionally-transformed scores will show very high correlations, even when their distributions are markedly different. . This was illustrated in the real data examples where, in spite of leading to diverging conclusions about the presence and strength of moderation effects, the three types of score were correlated with one another at >.97.

Although using IRT scores is more time consuming and technically demanding than using normalising transformation, it may be worth the additional effort. In addition to performing better in the current study, IRT scores can be estimated reasonably easily in a range of freely available software packages and have several practical and theoretical advantages over transformed sum scores. First, they are easily estimable in the presence of missing item data, or when respondents did not complete an identical set of items (Embretson & Reise, 2000). Second, the diversity of available IRT models means that many kinds of response formats, scale structures, or theories about how the latent trait relates to item responses can be accommodated. For example, a bi-factor model could be fit when it is desirable to partition general and specific trait variance captured by a set of items (Cai, Yang, & Henson, 2011); if a scale has a categorical response format, a nominal response model could be fit (Bock, 1972); or if items follow an ideal point process an unfolding model can be fit (e.g. Chernyshenko, Stark, Drasgow, & Roberts, 2007). All of these and other features can be, easily dealt with in an IRT framework, while posing significant problems or being simply impossible to take account of when using sum scores, both raw and transformed to normality.

Furthermore, while an IRT model can be chosen based on theoretical considerations, the choice of a transformation is somewhat arbitrary and usually driven by pragmatic

considerations. The choice of an IRT model can be evaluated both overall and with respect to individual items using well-studied goodness-of-fit statistics and graphical checks. A beneficial side effect of this is that the process of fitting and evaluating IRT model(s) is likely to encourage explicit consideration of the assumptions that underpin the phenotypic proxy used. However, no analogous tests exist for transformations. More importantly, from a conceptual perspective, if the genetic and environmental influences on the phenotype in the absence of the influence of the moderator are normally distributed and there is true GxE in the population then the phenotype *should* show a non-normal distribution because GxE involves an expansion (or contraction) of the variance in a phenotype according to the levels of moderator. This expansion (or contraction) of variance shows up in the marginal distribution of the phenotype as non-normality that is commensurate with the GxE effect. Using a transformation to normality is, therefore, directly at odds with theoretical expectations when GxE is hypothesised. For this reason, methodologies such as Box-Cox transformations which can optimise the normalisation of a distribution may actually perform worse than cruder methods such as log-transformations. The latter will almost always yield a worse approximation to normality but this worse approximation may retain some of the non-normality due to the interaction when one is present. In IRT models, the need to retain any non-normality that is genuinely due to a GxE is also a problem to some extent; however, the assumption of a normal latent distribution is not a necessity; where appropriate alternative prior distributions can be specified in a manner that is far more flexible than attempting to obtain that distribution through transformation of observed scores. Where both approaches are limited is that the underlying liability distribution absent the influence of the moderator could be non-normal due to other moderators or the effects of rare but highly influential etiological factors that engender extreme effects. Analogous to the problem of distinguishing non-normality due to moderation versus poor scaling, it is not easy to disentangle non-

normality due to the effect of a moderator of interest and non-normality due to other etiological factors without detailed a priori knowledge.

Further, the favourable performance of the IRT scores in the simulation study should be interpreted in light of the fact that they were estimated under idealised conditions. In practice their use is more complicated and may be less effective. For example, a graded response model was fit to the data because it was known that this model had been used to generate the item responses. Thus, there was no risk of mis-specifying the psychometric model. In reality, the appropriate model for the items will not be known in advance; it will have to be chosen on the basis of the item format and a hypothesis about how the latent trait is related to item responding and then tested for appropriateness. The lack of a priori knowledge about the appropriate IRT model for a given set of items increases the risk that the chosen model will be mis-specified in some important way. Further, parametric IRT models are also often poor fits to the very same kinds of data that prove problematic in GxE tests, such as those concerning psychopathological phenotypes. Less restrictive non-parametric IRT models are sometimes recommended as alternatives (Meijer & Baneke, 2004) but these methods do not allow estimation of factor scores for use in GxE tests. Finally, at a very pragmatic level, IRT models are only useful when item-level data are available, which is not always the case.

Another practical consideration when using IRT scores in tests of GxE, is the importance of assessing the empirical reliability of factor scores from IRT models, as one would for sum scores (see Culpepper, 2013). Unreliable IRT scores will not only be ineffective in addressing bias in GxE; they will also result in attenuated estimates of twin correlations and bias other model parameters (van den Berg et al., 2007). Similarly, as the extent to which the accuracy of the scores as measures of the intended underlying dimension depends on the appropriateness of the IRT model, its specification should be carefully

considered and its fit assessed empirically (see Embretson & Reise, 2013), bearing in mind that even the best fitting models will represent only an approximation to reality.

In practice, it is also worthwhile to compare results obtained using IRT scores with those obtained using raw and transformed sum scores. Comparison can highlight how sensitive results are to phenotypic scaling. Under some conditions, e.g. when the phenotype and moderator do not have strong associations or the phenotypic distribution departs only slightly from its population distribution, scaling of the phenotype may make little difference to results. In addition, in rare cases where the phenotypic distribution is mis-specified in the IRT model used to estimate the scores but well approximated by the sum scores, the sum scores could, in principle, produce less biased results than the IRT scores. Even when the phenotypic distribution is correctly assumed to be normal, no non-linear transformation or IRT score estimation method guarantees a perfect reconstruction of the phenotypic distribution as it exists in the population. In fact, as argued above, the scores produced by a transformation to normality could be 'too normal' in the sense that in the presence of GxE non-normality of the phenotype would usually be expected.

In sum, there are significant challenges in correctly choosing between a 'GxE' and a 'scaling' explanation for apparent GxE effects but the importance of doing so is considerable. For example, if a GxE explanation is incorrectly accepted over a scaling explanation, this falsely supports the view that some features of a person or their environment either constrains or supports the expression of genetic influences on some phenotype of interest. At an academic level, this can lead to theories which lack parsimony and which when further pursued may lead to wasted research efforts. Further, spurious GxE evidence can also falsely bolster the impression that a candidate moderating variable is an as important causal factor or potentially fruitful target of intervention. Thus, continued efforts should be invested into establishing appropriate scales for phenotypes submitted to GxE analyses.

Limitations

A limitation of the current study is that I did not directly compare the two-step IRT approach with a one-step approach presented here. A one-step approach has yet to be developed for testing of GxE within the Purcell's (2002) framework; however, it is possible to anticipate some of its disadvantages and advantages. First, the approach would share the limitation of the two-step approach that the true phenotypic distribution would not be known but assumed. Assuming a normal distribution for the phenotype when the true distribution is non-normal could, in principle, result in biased GxE tests in a similar way to using a poorly scaled sum score. It would also share the necessity to select an appropriate IRT model and freely estimate its parameters in a finite sample. A further disadvantage would be its statistical and computational complexity as compared to a two-step approach. However, an important advantage would be that the error-free latent trait could be decomposed directly and this is likely to result in less biased GxE tests.

Further, and perhaps most importantly, a one-step approach is more appropriate from a conceptual perspective because it provides a much more direct operationalization of GxE hypotheses. In the two-step approach, a distribution for the phenotype is assumed in the first step; however, in tests of GxE it is important to distinguish between assumptions about the marginal distribution of the phenotype and the distribution of the underlying genetic and environmental influences absent the influence of the moderator. While the former would be expected to be non-normal because being subject to moderation skews the phenotypic distribution, the latter can usually be assumed normal. The two-step approach unfortunately conflates these distinct contributions because it specifies a distribution only for the latent phenotype. In addition, although I designed the simulation conditions to be as realistic as possible, I covered only a limited range of the possible conditions that could occur in the real world. Although the principles discussed are likely general, I conducted analyses within

specific GxE and IRT frameworks and used a limited range of parameter values. Similarly, while inclusion of a real data example is important to test conclusions from simulation studies in a more ecologically valid context, these too are limited by their specificity.

Conclusions

Tests of GxE can be biased by inappropriate scaling of a phenotype, and reliance on raw scores that are significantly skewed is to be discouraged. Two potentially useful solutions are to transform sum scores to normality or to estimate IRT scores based on an appropriate model. Although these strategies will suffer low statistical power, they reduce the rate of spurious GxE detection and recover the correct direction of effects. Therefore, researchers can be more confident about the presence and direction of GxE when it is identified using one of these strategies than when using a raw sum score.

Footnotes[1]

Purcell's GxM approach requires assumption of a normal distribution for the phenotype conditional on the moderator; however, the presence of moderation will result in a skewed marginal distribution for the phenotype.

Chapter 5: Discussion

In this thesis, I have defined construct truncation as under-representation of the extremes of a variable in a research sample. In the preceding chapters I outlined ways in which construct truncation commonly arises and showed that its occurrence may be associated with serious distortions of the theoretical conclusions drawn from affected datasets. In this final chapter, I summarise and integrate the results of these chapters, highlight their limitations, and suggest future directions for research.

In **Chapter 2,** I focussed on construct truncation due to person selection, highlighting one area for which I believe the issue to be especially pertinent. I argued that studies involving psychopathological phenotypes are particularly vulnerable to the effects of range-restricting person selection due to frequent focus on clinically ascertained samples. Using an example from autism spectrum disorder (ASD) research, I presented a statistical model of range-restriction due to clinical diagnosis, assuming an underlying multivariate normal distribution of autistic traits in the population. Using this model, I evaluated the extent to which estimates of associations between symptom domains in ASD are under-estimated in clinically ascertained samples. Results suggested that the downward bias in estimates could be substantial, especially when considering associations involving the restricted repetitive activities symptom of the classical triad of ASD. A real data example also demonstrated that associations between ASD traits were much smaller when analyses were restricted to clinically diagnosed individuals.

In **Chapter 3,** I considered a more subtle example of construct truncation due to person selection, highlighting the ease with which it can arise undetected. I considered, in detail, a specific example of possible construct truncation from individual differences

research: the ‘intelligence compensation hypothesis’ (ICH). The ICH is that cognitive ability and conscientiousness are negatively correlated due to a tendency for individuals to calibrate their effort levels to ability levels. I hypothesised that previous evidence for the ICH is artifactual and has resulted from construct truncation on achievement in many samples. I argued that the negative associations between cognitive ability and conscientiousness cited in support of ICH have been due to the fact that the samples in which these associations were observed tended to include only individuals who showed certain levels of occupational or educational accomplishment. Taking a sample with relatively little such non-random selection I found that the associations between cognitive ability and conscientiousness were positive. However, artificially introducing selection on achievement resulted in attenuation of these positive estimates and in half the cases reversals in their direction. Together, this evidence suggests not only that intelligence and conscientiousness are probably not generally negatively correlated in the full population, but often become so when samples are selected based on specified levels of achievement. Estimates of their association may also be substantively negatively biased when selection on achievement is indirect. For example, despite researcher’s intentions, it is common that their samples are more educated and thus likely also of higher average IQ and conscientiousness than the general population.

Finally, in **Chapter 4**, I considered the problem construct truncation due to item selection. I developed a model of construct-truncating item selection to study its effects on gene-environment interactions, as well as compare possible solutions. These simulations suggested that using raw sum scores as measures of the phenotype of interest can produce substantial bias in estimates of gene-environment interaction; however, this bias can be substantially mitigated by transforming the raw sum scores to normality or by using IRT scores. Two real data examples showed that the choice of score: raw sum, transformed sum, or IRT score, can lead to quite different conclusions. In one example, using a raw sum score

suggested statistically significant GxE while an IRT score suggested no GxE. Together, these results suggested that construct-truncating item selection has the potential to and may already have misled researchers interested in evaluating GxE interactions. Fortunately, two simple-to-use fixes showed promise in mitigating these effects, namely transformation of raw scores and use of IRT scores.

A common message across all chapters is that there are commonly occurring circumstances under which construct truncation can have important implications for the accuracy of substantive inferences made based on affected datasets. Our increasingly sophisticated theories about interplay among variables and characterisation of constructs whose tests require use of increasingly complex statistical methodologies are especially susceptible; for example, construct truncation can even reverse the direction of apparent GxE interaction effect whilst barely affecting the association between phenotype and moderator. Compounding the problem, the presence of construct truncation may not be obvious as in the example of the ICH in which the truncation occurred primarily on a variable that was not explicitly measured. In the following section I, therefore, make some recommendations for detecting, quantifying and mitigating the effects of construct truncation.

It would be easy to recommend that research studies avoid problems of construct truncation by utilising samples and measures that reliably capture the full range of variability in their constructs of interest. Unfortunately, this is extremely difficult to implement in practice. As noted in the introduction, construct truncation due to person selection is often beyond complete control of the researcher because active participation in the vast majority of studies is voluntary, and individuals often vary systematically in interest and motivation to participate (e.g. Marcus & Schütz, 2005). The same is generally true when participation is passive, as, for example, when banks of anonymised data gathered for other purposes, such as American Scholastic Assessment Test scores, are used (e.g. Hunt & Madyastha, 2008; Lee &

Valliant, 2009). Likewise, construct truncation due to item selection may be difficult to avoid because of the need to rely on validated questionnaires which are unlikely to have been developed with explicit focus on their *ranges* of reliable measurement (Thomas, 2011). Thus, my recommendations aim to acknowledge these difficulties and suggest practical steps that can be taken to minimise problems of construct truncation.

Recommendations concerning construct truncation due to person selection

Minimising construct truncation begins with recruitment of participants. The goal is to recruit participants who are representative of a specified target population. Certain recruitment strategies have been associated with better sample representativeness. These include providing specific training for the project workers responsible for recruitment; investigating and addressing reasons for non-participation; providing appropriate incentives or reassurances to overcome barriers to non-participation; sending multiple reminders; and attempting to contact target participants via multiple means including phone, personalised letters, and door-knocking (where appropriate); and enriched recruitment in vulnerable subgroups. These strategies have been shown to increase participation rates especially amongst individuals 'at risk' of non-participation and whose presence in the sample is crucial to its representativeness (Eisner & Ribeaud, 2007). However, they are also resource- and time- intensive, and for many studies will be unrealistic with limited resources.

Given the practical difficulties of recruiting a representative sample, it is always advisable to consider the possibility that a sample has been subject to construct truncation, even when there has been no explicit exclusion of individuals with more extreme trait levels. The sampling strategy should be evaluated conceptually and an attempt made to understand the extent to which it was likely to have disproportionately missed individuals at high or low levels of the relevant constructs. For example, one of the most commonly used convenience

samples is that of university or college students. As noted by several authors, researcher using such a sample should consider the fact that they tend to exclude individuals of lower cognitive ability due to the standards of academic achievement that have to be reached to enter the population of university or college students (e.g. Chamorro-Premuzic & Furnham, 2004; Hägglund & Larsson, 2006). More generally, for any construct with negative social connotations, it is typically the individuals who are highest on that construct that are the most difficult to recruit and retain. It is the individuals with the most problematic behaviour that are least likely to participate in studies of crime (Eisner & Ribeaud, 2007), individuals with the highest level of psychopathology who are least likely to respond in psychiatric epidemiology studies (e.g. Kessler et al,. 2005; Merikangas et al., 2010), and individuals with the lowest levels of cognitive ability and greatest decline in it who are least likely to participate in studies of cognitive decline (Deary et al., 2011). There are also some variables that are commonly measured and associated with participation in research studies irrespective of the nature of the study. These may serve as red flags for construct truncation when seen to diverge from their expected distributions. Demographic variables associated with being less likely to participate in research include low levels of education, male sex, being a member of an ethnic minority, poor physical and/or mental health, and low socioeconomic status (Bechger et al., 2002; van Goor et al., 2005; Volken, 2013).

Post-data collection, to evaluate the presence and gauge the degree of construct truncation due to person selection, sample means and variances can sometimes be compared with normative means and variances (e.g. Costa, McCrae, Zonderman, Barbano, Lebowitz & Larson, 1986). Here, substantial departures of sample from normative summary statistics would indicate cause for concern. Availability of normative data is, however, the exception rather than the rule and even where available is likely to have itself been subject to some construct truncation due to person selection in the norming sample because of reliance on

volunteers (Marcus & Schütz, 2005; Vink et al., 2004). Comparison of sample to normative data can also sometimes be misleading because attending to reductions in the variances of observed variables without considering the selection mechanism underlying those reductions could lead to false conclusions about selection effects. One example would be to conclude that selection effects are minimal when phenotypic variance in a measure is the same as in a normative sample when in fact a decrease in the systematic variance in that measure was offset by an increase in error variance. This could occur, for example, if a test is administered to a sample for whom the reading level of the test exceeds the reading ability of many respondents. Similarly, comparison of sample to normative data requires the assumption that no construct truncation due to item selection has occurred. If it has, the normative data will not reliably capture the range of the construct as it occurs in the target population.

In the absence of normative data, efforts to obtain some basic information from individuals who did not participate can also provide valuable information about whether construct truncation is likely to have been a problem and may even allow corrections for construct truncation to be made (Vink et al., 2004). Where this is not possible, examining the factors associated with responding later or with dropping out may be informative about the factors associated with complete non-participation. It is not, however, necessarily valid to consider late responders/dropouts and non-responders as though they were simply at different points on the same continuum of participation because of qualitative differences in reasons for non-response (Studer et al., 2013). Those who participate late, for example, may simply need additional facilitation, reminders or incentives to participate whereas complete non-responders may object to the study or to the notion of handing information about themselves over. The availability and fidelity of information about the composition of the population as a whole or of the non-participating subsection is critical here. It is not, for example, sufficient to examine response rates because there is no guarantee of a straightforward relation between

bias due to non-response and extent of non-response (Stang, 2003). Non-response may be high but unrelated to the construct of interest. On the other hand it may be low except for specific sociocultural groups with more extreme average levels of the construct of interest (Eisner & Ribeaud, 2007).

It may be possible to apply statistical corrections to the data or parameter estimates with the aim of obtaining an unbiased estimate of a parameter affected by construct truncation. For example, depending on the information available in a given study and the assumptions that can reasonably be made about the distribution of that construct in the target population, possible strategies include selection models, data weighting, range-restriction corrections, or censored (e.g. tobit) regression (e.g. Asparouhov, 2005; Kamakura & Wedel, 2001; Nie, Chu, & Korostyshevskiy, 2008; Sackett & Yang, 2000). These strategies have some important limitations which boil down to the fact that unless information about the selection mechanism and/or the target population is known (or can be reasonably inferred), these corrections will not yield unbiased estimates. For example, in range restriction corrections for Pearson correlations, there are many ways in which bias can result including: selecting the wrong formula for a given selection scenario, violation of the assumptions of linearity and homoscedasticity of the regression of the selection variable(s) on the construct of interest, assigning the wrong roles to variables; or submitting incorrect estimates of the degree of range restriction or population variances of variables (Alexander et al., 1984; Linn, 1983; Schmidt, Oh, & Le, 2006; Sackett et al., 2007). Thus, the best approach overall may be to estimate upper and lower bounds of the effect using corrected and uncorrected estimates together with a range of plausible assumptions about the selection mechanism and/or distribution of the construct in the population.

Recommendations concerning construct truncation due to item selection

Minimising construct truncation due to item selection begins at the test development stage. Traditionally, test developers have sought to create tests with maximal internal consistency; however, in doing so they may have inadvertently restricted thereliable range of measurement. This is because in aiming to maximise internal consistency, items are generally selected to form a highly correlated set. Highly correlated items will tend to have very similar response distributions, implying that they tap similar levels of a construct. Furthermore, selecting items to maximise their internal consistency provides no guarantee of high test-retest reliability and may in fact undermine the validity of a scale if it results in the omission of item covering important content areas (e.g. McCrae, Kurtz, Yamagata & Terracciano, 2011). To both avoid restriction of content breadth and to ensure the measurement of an appropriate range of construct levels with good precision, items should aim to capture a wide range of construct levels, even if this sacrifices internal consistency to some degree. For example, developers of measures of aggression should consider including not only behaviours indicative of high levels of overt aggression (e.g. hitting, kicking, physical conflicts etc.), but also low to middling levels (e.g. shyness, assertativeness, suppressed anger, angry ideations; Anholt & Mackay, 2012).

One area where a restricted range of reliable measurement is of particular importance is in measuring psychopathological constructs. Here, additional factors play into the selection of items which tap limited ranges of trait levels. First, many items in many psychopathological scales have been selected based on their ability to discriminate between diagnosed cases and non-cases. These items will tend to be very good at measuring trait levels at and near a clinical diagnostic cut-off point but likely at the expense of reliably measuring other trait levels. Second, items tapping trait levels at and above a diagnostic cut-off point tend to be selected because they have the greatest face validity. Conceptualising sub-clinical psychopathological trait levels is relatively new, and researchers may have less

previous research to inform the writing of items tapping these levels. This makes it difficult to test the very idea that psychopathologies are the extremes of common trait levels empirically too.

Perhaps the most promising approach to addressing construct truncation due to item selection is fitting parametric item response theory models to estimate item and test information across the range of trait values during test development. Parametric IRT models, unlike the majority of other test development and evaluation methods, acknowledge that the precision of measurement is not equal across the entire range of construct values. Computing the test information function allows evaluation of the locations along the presumed latent trait continuum that are relatively more and less precisely measured. Construct truncation is in evidence when a test cannot measure a trait at one or both extremes with adequate measurement precision. Estimating item difficulties can identify the regions of the continuum that specific items are capable of most reliably measuring. On identifying regions of low information that lack in items at all, or where items have poor discrimination, items may be written or modified to extend the range of reliable measurement of the test above and/or below its existing range. However, it is important to acknowledge there will be limits to the number of items that can be administered in a given study, such that including many items to achieve a favourable reliable range of measurement may come at the expense of other test properties. For example, there may be a need to weigh reliable range of measurement against the conceptual breadth of the construct that can be measured: an instantiation of the 'bandwidth-fidelity dilemma' (e.g. Ones & Viswesvaran, 1996). However, researchers should also consider the various available methods of mitigating this trade-off such as computerised adaptive testing (e.g. Pilkonis, Choi, Reise, Stover, Riley & Cella, 2011) or planned missingness designs (e.g. Rhemtulla & Little, 2012). The former minimises the number of items administered by way of an algorithm that selects items that are expected to be most

informative about a respondent's trait level. The latter can be used to reduce the number of items administered through strategic omission of some items and later correction for missingness.

In Chapter 4, I also showed that, provided the IRT model is correctly specified, factor scores estimated from it will provide a good estimate of a true GxE effect in most cases. This is consistent with other previous research showing bias reductions in other moderation models (e.g. Kang & Waller, 2005; Morse et al., 2012). It may even be possible to use this technique to undo some of the effects of construct truncation due to person selection although this remains to be determined.

Future Directions

Across the previous chapters, I noted that there are statistical methods available for correcting data or parameters for construct truncation provided some information is known or can be reasonably assumed. Their limitation is, in particular, in the availability of information about the distribution (especially variance or shape) of a construct in the population. Thus, increasing our knowledge of these underlying distributions is an important area for future research. As one example, use of IRT models in the case of psychopathological phenotypes requires an assumption to be made about their underlying distribution in the population. I took the approach of assuming that a normal distribution characterised the distributions of these populations. This is consistent with much of current opinion in psychopathology research which assumes that psychopathological traits are merely the upper extremes of normal distributions, not qualitatively distinct states (Caspi et al., 2014). The rationale for this is based on various pieces of evidence: the highly polygenic nature of psychopathological traits, observed normal distributions of psychopathological traits in the general population, presence of sub-clinical levels of psychopathological traits in relatives of individuals with a

clinical diagnosis, and movement of individuals into and out of the clinical range of psychopathological traits over the course of their life (e.g. Baron-Cohen et al., 2001; Cichetti & Rogosch, 2002; Wray et al., 2014).

Indicative of their degree of acceptance within the research community, models of psychopathological traits that assume underlying continua are providing a basis for the development of a research classification system for psychopathological disorders (Cuthbert & Insel, 2013). These kinds of models are also already being built on, adapted and extended to answer questions about psychopathological etiology such as group differences in prevalence. For example, for phenotypes such as ASD, attention-deficit hyperactivity disorder (ADHD) and aggression in which males are disproportionately affected, multi-factorial threshold theory states that a continuous (usually assumed normal) etiological liability distribution underpins psychopathological phenotypes; however, males have a lower threshold for manifesting clinical levels of the trait (Hamshere et al., 2013; Lai et al., 2015; Tuvblad et al., 2006). That is, females with same 'etiological load' as males would be less likely to qualify for clinical diagnosis because they are in some way more protected against exhibiting the maladaptive behaviours on which clinical diagnosis is based.

Certainly, some conditions show a liability distribution that would be expected to be non-normal: Alzheimer's disease is, for example, known to be influenced by at least one allele of disproportionately large effect: the APOE e4 allele (Genin et al., 2011). Even in phenotypes in which a highly polygenic model is generally accepted, arguments can also be made for underlying non-normal liability distributions due to processes such as GxE, and intra- and inter-allelic interaction or the effects of powerful causal (genetic or environmental) factors (e.g. see van den Oord et al., 2003). As discussed in chapter 4, the presence of GxEs will tend to expand the variance in a trait at one end of its distribution, equivalent to introducing skewness. Recent replication crises notwithstanding, the very large number of

published GxE studies published to date would suggest at least a perception in the field that GxEs must be common (Dick et al., 2015).

Some authors have argued that there may not be meaningful variation in psychopathological traits below clinical cut-off points at all; that these traits are better characterised as 'quasi-traits' (Reise & Waller, 2009). Overall, however, it would seem most plausible to consider the distributions of most psychopathological traits to be mixture distributions. In intellectual disability, for example, many idiopathic cases may simply represent the lower extreme of a continuous distribution; however, many others (e.g. Down syndrome) clearly represent the effects of a single, powerful genetic or environmental insult.

Some studies have aimed to identify the appropriate distributions to characterise psychopathological traits by comparing the fit of various models assuming different distributions (e.g. van den Oord et al., 2003). However, the question of the appropriate distribution to assume for a given phenotype is unlikely to be answered solely by examining the fit of various distributions to the underlying latent distributions of these traits for reasons discussed in Chapter 4 relating to the fact that a multitude of statistically indistinguishable states could underlie the same observed data. Rather, continuing to make progress in understanding the etiology of complex traits is likely to be critical because this can directly inform on the reasonableness of distributional assumptions. Although it was once necessary (at least in practical terms) to assume multivariate normality for parameter estimation, recent and continuing developments in statistical methodology, especially availability of Bayesian estimation techniques which make complex models more tractable, mean that this is no longer always the case. Thus, the primary limiting factor is likely theoretical knowledge to inform the distributional shape to assume, rather than the statistical models to operationalise it.

Limitations

The primary limitations of this thesis concern the extent to which the simulation studies captured the range of circumstances likely to be occurring in the real world. While I intended the simulation designs across the chapters to cover a range of plausible real world conditions, there are inevitable limits to the range of variables that can be manipulated in any given study. In addition, I considered construct truncation due to person and item separately, whereas in the real world they are likely to co-occur and interact. In fact, this is one way in which construct truncation may go undetected: if a measure with a limited range of reliable measurement is administered to a sample with a correspondingly limited range of trait values then there will be few clues in the data that construct truncation has occurred. I also did not discuss the important issue of how to determine what the target population should be. Whether or not construct truncation is relevant depends on whether the aim is to generalise to a target population that exhibits only a limited range of possible construct levels. However, it is not always easy to determine whether it is appropriate to consider this kind of restricted population rather than a more general population as, for example, when it is not clear if clinical levels of a trait really do have the same meaning and origin as sub-clinical trait levels. Further, even if considered appropriate, analysing a restricted population is likely to entail issues such as violations of normality due to dichotomising of a continuous distribution.

Finally, I did not address in any detail the practical challenges of developing questionnaires that have the same meaning and measurement properties at high and low levels of a construct (e.g. above and below clinical cut-off points; Murray et al., 2014). This is particularly a problem for constructs in which questionnaire responding may be directly related to the trait of interest. For example, individuals who are high in neuroticism are more likely to use middle response options, thus resulting in systematic underestimation of their trait levels (Murray, Molenaar, & Booth, submitted).

Conclusions

Construct truncation can have important implications for the theoretical inferences made in empirical research. Statistical solutions can help to mitigate its effects; however, they are limited by the need to make assumptions about the distribution of the construct in the population and/or the selection mechanism that intervened between population and sample. In contributing to resolving this limitation, improved understandings of the population distributions and underlying etiologies of specific phenotypes are likely to be critical in identifying, quantifying and correcting for construct truncation when it occurs.

References

Aitken, A.C. (1935). Note on selection from a multivariate normal population. *Proceedings of the Edinburgh Mathematical Society (Series 2)*, *4*, 106-110.Austin, E.J. (2005). Personality correlates of the broader autism phenotype as assessed by the Autism Spectrum Quotient (AQ). *Personality and Individual Differences*, *38*, 451-460.

Aguinis, H. (1995). Statistical power with moderated multiple regression in management research. *Journal of Management*, *21*, 1141-1158.

Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology*, *82*, 192.

Alexander, R. A., Carson, K. P., Alliger, G. M., & Barrett, G. V. (1984). Correction for restriction of range when both X and Y are truncated. *Applied Psychological Measurement*, *8*, 231-241.

Alexander, R.A., Carson, K.P., Alliger, G.M., & Cronshaw, S.F. (1989). Empirical distributions of range restricted SD x in validity studies. *Journal of Applied Psychology*, *74* 253-258.

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "red flags" for autism screening: the short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*, 202-212.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders (4th edition).* Washington, DC: American Psychiatric Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.).* Washington, DC: *American Psychiatric Association*.

Anholt, R. R., & Mackay, T. F. (2012). Genetics of aggression. *Annual Review of Genetics, 46*, 145-164.

Asbury, K., Wachs, T. D., & Plomin, R. (2005). Environmental moderators of genetic influence on verbal and nonverbal abilities in early childhood. *Intelligence, 33*, 643-661.

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12*, 411-434.

Baird, G., Simonoff, E., Pickles, A., Chandler, S., Loucas, T., Meldrum, D., et al. (2006). Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: the Special Needs and Autism Project (SNAP). *The Lancet, 368*, 210-215.

Baker, T. J., & Bichsel, J. (2006). Personality predictors of intelligence: Differences between young and cognitively healthy older adults. *Personality and Individual Differences, 41*, 861-871.

Bartels, M., van Weegen, F. I., van Beijsterveldt, C. E., Carlier, M., Polderman, T. J., Hoekstra, R. A., & Boomsma, D. I. (2012). The five factor model of personality and intelligence: A twin study on the relationship between the two constructs. *Personality and Individual Differences, 53*, 368-373.

Baron-Cohen, S., Scott, F.J., Allison, C., Williams, J., Bolton, P., Matthews, F.E., et al. (2009). Prevalence of autism-spectrum conditions: UK school-based population study. *The British Journal of Psychiatry, 194*, 500-509.

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders, 34*, 163-175.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*, 5-17.

Beasley, T. M., Erickson, S., & Allison, D. B. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited?. *Behavior Genetics, 39*, 580-595.

Bechger, T. M., Boomsma, D. I., & Koning, H. (2002). A limited dependent variable model for heritability estimation with non-random ascertained samples. *Behavior Genetics, 32*, 145-151.

Biederman, J., Kwon, A., Aleardi, M., Chouinard, V. A., Marino, T., Cole, H., ... & Faraone, S. V. (2005). Absence of gender effects on attention deficit hyperactivity disorder: findings in nonreferred subjects. *American Journal of Psychiatry*, *162*, 1083-1089.

Berry, C.M., Clark, M.A., & McClure, T.K. (2011). Racial/ethnic differences in the criterion-related validity of cognitive ability tests: A qualitative and quantitative review. *Journal of Applied Psychology, 96*, 881-906.

Bishara, A. J., & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods*, *17*, 399-417.

Bobko, P., Roth, P. L., & Bobko, C. (2001). Correcting the effect size of d for range restriction and unreliability. *Organizational Research Methods, 4,* 46-61.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bogg, T., & Roberts, B. W. (2013). The case for conscientiousness: Evidence and implications for a personality trait marker of health and longevity. *Annals of Behavioral Medicine, 45,* 278-288.

Boomsma D. I., Martin N. G. (2002) Gene–environment interactions. In: D'haenen H., den Boer J.A., Willner P (eds) *Biological Psychiatry*. Wiley, New York, pp 181–187.

Booth, T., Murray, A.L., McKenzie, K., Kuenssberg, R., O'Donnell, M., & Burnett, H. (2013). Brief Report: An evaluation of the AQ-10 as a brief screening instrument for ASD in adults. *Journal of Autism and Developmental Disorders, 43,* 2997-3000.

Bronfenbrenner, U., & Ceci, S. J. (1994). Nature-nuture reconceptualized in developmental perspective: A bioecological model. *Psychological Review, 101,* 568- 586.

Brunsdon, V.E., & Happé, F. (2014). Exploring the 'fractionation'of autism at the cognitive level. *Autism, 18,* 17-30.

Burt, S. A., & Klump, K. L. (2009). The etiological moderation of aggressive and nonaggressive antisocial behavior by age. *Twin Research and Human Genetics, 12,* 343-350.

Button, T. M., Hewitt, J. K., Rhee, S. H., Corley, R. P., & Stallings, M. C. (2010). The moderating effect of religiosity on the genetic variance of problem alcohol use. *Alcoholism: Clinical and Experimental Research, 34,* 1619-1624.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16,* 221-248.

Carter, A. S., Black, D. O., Tewani, S., Connolly, C. E., Kadlec, M. B., & Tager-Flusberg, H. (2007). Sex differences in toddlers with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 37,* 86-97.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales. *Journal of Personality and Social Psychology, 67,* 319-333.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., ... & Moffitt, T. E. (2014). The p factor: one general psychopathology factor in the structure of psychiatric disorders?. *Clinical Psychological Science,2,* 119-137.

Chalmers, R.P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48,* 1-29.

Chamorro-Premuzic, T., & Furnham, A. (2004). A possible model for understanding the personality-intelligence interface. *British Journal of Psychology, 95,* 249-264.

Chamorro-Premuzic, T., Moutafi, J., & Furnham, A. (2005). The relationship between personality traits, subjectively-assessed and fluid intelligence. *Personality and Individual Differences, 38,* 1517-1528.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: toward increasing the flexibility of personality measures. *Psychological Assessment, 19,* 88-106.

Cicchetti, D., & Rogosch, F. A. (2002). A developmental psychopathology perspective on adolescence. *Journal of Consulting and Clinical Psychology,70,* 6-20.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Crawford, J. R., & Henry, J. D. (2003). The Depression Anxiety Stress Scales (DASS): Normative data and latent structure in a large non-clinical sample. *British Journal of Clinical Psychology*, *42,* 111-131.

Constantino, J.N., Gruber, C.P., Davis, S., Hayes, S., Passanante, N., & Przybeck, T. (2004). The factor structure of autistic traits. *Journal of Child Psychology and Psychiatry*, *45,* 719-726.

Costa Jr, P. T., McCrae, R. R., Zonderman, A. B., Barbano, H. E., Lebowitz, B., & Larson, D. M. (1986). Cross-sectional studies of personality in a national sample: II. Stability in neuroticism, extraversion, and openness. *Psychology and Aging*, *1,* 144.

Culpepper, S. A. (2013). The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution. *Applied Psychological Measurement, 37,* 201-225.

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Medicine*, *11*, 126.

Deary, I. J., Gow, A. J., Pattie, A., & Starr, J. M. (2012). Cohort profile: the Lothian Birth Cohorts of 1921 and 1936. *International Journal of Epidemiology*, *41,* 1576-1584.

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of marketing research, 38*, 269-277.

Dick, D. M., Agrawal, A., Keller, M. C., Adkins, A., Aliev, F., Monroe, S., ... & Sher, K. J. (2015). Candidate Gene–Environment Interaction Research Reflections and Recommendations. *Perspectives on Psychological Science,10*, 37-59.

Distel, M. A., Middeldorp, C. M., Trull, T. J., Derom, C. A., Willemsen, G., & Boomsma, D. I. (2011). Life events and borderline personality features: the influence of gene–environment interaction and gene–environment correlation. *Psychological Medicine, 41*, 849-860.

Dollinger, S. J., & Leong, F. T. (1993). Volunteer bias and the five-factor model. *The Journal of Psychology, 127*, 29-36.

Dominicus, A., Palmgren, J., & Pedersen, N. L. (2006). Bias in variance components due to nonresponse in twin studies. *Twin Research and Human Genetics, 9*, 185-193.

Donders, J., Elzinga, B., Kuipers, D., Helder, E., & Crawford, J. R. (2013). Development of an eight-subtest short form of the WISC-IV and evaluation of its clinical utility in children with traumatic brain injury. *Child Neuropsychology,19*, 662-670.

Dworzynski, K., Ronald, A., Bolton, P., & Happé, F. (2012). How different are girls and boys above and below the diagnostic threshold for autism spectrum disorders?. *Journal of the American Academy of Child & Adolescent Psychiatry, 51*, 788-797.

Dykiert, D., Gale, C. R., & Deary, I. J. (2009). Are apparent sex differences in mean IQ scores created in part by sample restriction and increased male variance?. *Intelligence, 37*, 42-47.

Eaves, L. J., Last, K., Martin, N. G., & Jinks, J. L. (1977). A progressive approach to non-additivity and genotype-environmental covariance in the analysis of human differences. *British Journal of Mathematical and Statistical Psychology, 30,* 1-42.

Eaves, L. J. (2006). Genotype× environment interaction in psychopathology: fact or artifact?. *Twin Research and Human Genetics,* 9, 1-8.

Eisner, M., & Ribeaud, D. (2007). Conducting a Criminological Survey in a Culturally Diverse Context Lessons from the Zurich Project on the Social Development of Children. *European Journal of Criminology*, *4*, 271-298.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement,20,* 201-212.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.

Etter, J. F., & Perneger, T. V. (1997). Analysis of non-response bias in a mailed health survey. *Journal of Clinical Epidemiology, 50,* 1123-1128.

Eurich, T.L., Krause, D.E., Cigularow, K., & Thorton III, G.C. (2009). Assessment Centers: Current Practices in the United States. *Journal of Business Psychology,* 24, 387-407.

Facon, B., Magis, D., & Belmont, J. M. (2011). Beyond matching on the mean in developmental disabilities research. *Research in Developmental Disabilities,32,* 2134-2147.

Falconer, D. S., & Mackay, T. F. (1996). Introduction to quantitative genetics. Harlow. UK: Longman.

Fife, D. A., Mendoza, J. L., & Terry, R. (2012). The assessment of reliability under range restriction: A Comparison of α, ω, and test–retest reliability for dichotomous data. *Educational and Psychological Measurement, 72*, 862-888.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*, 189-198.

Fombonne, E. (2003). Epidemiological surveys of autism and other pervasive developmental disorders: an update. *Journal of Autism and Developmental Disorders, 33*, 365-382.

Frazier, T. W., Youngstrom, E. A., Sinclair, L., Kubu, C. S., Law, P., Rezai, A., ... & Eng, C. (2010). Autism spectrum disorders as a qualitatively distinct category from typical behavior in a large, clinically ascertained sample. *Assessment, 17*, 308-320.

Frazier, T.W., Youngstrom, E.A., Speer, L., Embacher, R., Law, P., Constantino, J., et al. (2012). Validation of proposed DSM 5 criteria for Autism Spectrum Disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 51*, 28-40.

Furnham, A., Dissou, G., Sloan, P., & Chamorro-Premuzic, T. (2007). Personality and intelligence in business people: A study of two personality and two intelligence measures. *Journal of Business Psychology, 22*, 99-109.

Furnham, A., Moutafi, J., & Chamorro-Premuzic, T. (2005). Personality and intelligence: Gender, the Big Five, self-estimated and psychometric intelligence. *International Journal of Selection and Assessment, 13*, 11-24.

Gaughan, E.T., Miller, J.D., Pryor, L.R. & Lynam, D.R. (2009). Comparing Two Alternative Measures of General Personality in the Assessment of Psychopathology: A Test of the NEO PI-R and the MPQ. *Journal of Personality,* 77, 965-995.

Genia, V. (2001). Evaluation of the spiritual well-being scale in a sample of college students. *The International Journal for the Psychology of Religion,11,* 25-33.

Genin, E. Hannequin, D., Wallon. D., Sleegers, K., Hiltunen, M., Combarros, O.,… & Van Broeckhoven, C. (2011). APOE and Alzheimer disease: A major gene with semi-dominant inheritance. *Molecular Psychiatry, 16,* 903-907.

Ghiselli. E. E., Campbell, J. P., Zedeck, S. (1981). *Measurement and theory for the behavioural sciences.* San Francisco: WH Freeman & Co.

Gomez, R., Cooper, A., & Gomez, A. (2005). An item response theory analysis of the Carver and White (1994) BIS/BAS scales. *Personality and Individual Differences, 39,* 1093-1103.

Hägglund, G., & Larsson, R. (2006). Estimation of the correlation coefficient based on selected data. *Journal of Educational and Behavioral Statistics, 31,* 377-411.

Hamshere, M. L., Langley, K., Martin, J., Agha, S. S., Stergiakouli, E., Anney, R. J., ... & Thapar, A. (2014). High loading of polygenic risk for ADHD in children with comorbid aggression. *American Journal of Psychiatry*, 170, 909-916.

Hanges, P. J., Rentsch, J. R., Yusko, K. P., & Alexander, R. A. (1991). Determining the appropriate correction when the type of range restriction is unknown: Developing a sample-based procedure. *Educational and psychological measurement, 51,* 329-340.

Happé, F. & Ronald, A. (2008). The 'fractionable autism triad': a review of evidence from behavioural, genetic, cognitive and neural research. *Neuropsychology Review, 18,* 287-304.

Harden, K. P., Turkheimer, E., & Loehlin, J. C. (2007). Genotype by environment interaction in adolescents' cognitive aptitude. *Behavior Genetics,37,* 273-283.

Hartley, S. L., & Sikora, D. M. (2009). Sex differences in autism spectrum disorder: An examination of developmental functioning, autistic symptoms, and coexisting behavior problems in toddlers. *Journal of Autism and Developmental Disorders, 39,* 1715-1722.

Hays, R. D., Liu, H., Spritzer, K., & Cella, D. (2007). Item response theory analyses of physical functioning items in the medical outcomes study. *Medical Care, 45,* S32-S38.

Heath, A. C., Madden, P. A., & Martin, N. G. (1998). Assessing the effects of cooperation bias and attrition in behavioral genetic research using data-weighting. *Behavior Genetics, 28,* 415-427.

Hicks, B. M., DiRago, A. C., Iacono, W. G., & McGue, M. (2009). Gene–environment interplay in internalizing disorders: consistent findings across six environmental risk factors. *Journal of Child Psychology and Psychiatry, 50,* 1309-1317.

Hicks, B. M., South, S. C., DiRago, A. C., Iacono, W. G., & McGue, M. (2009). Environmental adversity and increasing genetic risk for externalizing disorders. *Archives of General Psychiatry, 66,* 640-648.

Hill, A., Roberts, J., Ewings, P., & Gunnell, D. (1997). Non-response bias in a lifestyle survey. *Journal of Public Health, 19,* 203-207.

Hoekstra, R.A., Bartels, M., Cath, D.C., & Boomsma, D.I. (2008). Factor structure, reliability and criterion validity of the Autism-Spectrum Quotient (AQ): A study in Dutch population and patient groups. *Journal of Autism and Developmental Disorders, 38*, 1555-1566.

Hoekstra, R.A., Vinkhuyzen, A.A., Wheelwright, S., Bartels, M., Boomsma, D.I., Baron-Cohen, S., et al. (2011). The construction and validation of an abridged version of the autism-spectrum quotient (AQ-Short). *Journal of Autism and Developmental Disorders, 41*, 589-596.

Hunt, E., & Madhyastha, T. (2008). Recruitment modeling: An analysis and an application to the study of male–female differences in intelligence. *Intelligence,36*, 653-663.

Hunter, J.E., Schmidt, F.L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91*, 594-612.

Iacono, W. G., Carlson, S. R., Taylor, J., Elkins, I. J., & McGue, M. (1999). Behavioral disinhibition and the development of substance-use disorders: Findings from the Minnesota Twin Family Study. *Development and Psychopathology, 11*, 869–900.

Johnson, W., & Krueger, R. F. (2005). Genetic effects on physical health: lower at higher income levels. *Behavior Genetics, 35*, 579-590.

Johnson, W., Kyvik, K. O., Mortensen, E. L., Skytthe, A., Batty, G. D., & Deary, I. J. (2011). Does education confer a culture of healthy behaviour? Smoking and drinking patterns in Danish twins. *American Journal of Epidemiology, 173*, 55-63.

Jones, D. N., & Paulhus, D. L. (2014). Introducing the Short Dark Triad (SD3) A Brief Measure of Dark Personality Traits. *Assessment, 21*, 28-41.

Kamakura, W. A., & Wedel, M. (2001). Exploratory Tobit factor analysis for multivariate censored data. *Multivariate Behavioral Research, 36*, 53-82.

Kang, S. M., & Waller, N. G. (2005). Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement, 29*, 87-105.

Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*, 617-627

Kolevzon, A., Smith, C.J., Schmeidler, J., Buxbaum, J.D., & Silverman, J.M. (2004). Familial symptom domains in monozygotic siblings with autism. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*,*129*, 76-81.

Kreiser, N. L., & White, S. W. (2014). ASD in females: Are we Overstating the gender difference in diagnosis?. *Clinical Child and Family Psychology Review*,*17*, 67-84.

Krueger, R. F., & Johnson, W. (2002). The Minnesota twin registry: current status and future directions. *Twin Research, 5*, 488-492.

Kuenssberg, R., Murray, A.L., Booth, T., & McKenzie, K. (2014). Structural validation of the abridged Autism Spectrum Quotient–Short Form in a clinical sample of people with autism spectrum disorders. *Autism*, *18*, 69-75.

Lai, M. C., Lombardo, M. V., Auyeung, B., Chakrabarti, B., & Baron-Cohen, S. (2015). Sex/gender differences and autism: setting the scene for future research. *Journal of the American Academy of Child & Adolescent Psychiatry*,*5*, 11-24.

Latvala, A., Dick, D. M., Tuulio-Henriksson, A., Suvisaari, J., Viken, R. J., Rose, R. J., & Kaprio, J. (2011). Genetic correlation and gene–environment interaction between alcohol problems and educational level in young adulthood. *Journal of Studies on Alcohol and Drugs, 72,* 210-220.

Lawley, D.N. (1944). A Note on Karl Pearson's selection formulæ. *Proceedings of the Royal Society of Edinburgh. Section A. Mathematical and Physical Sciences, 62,* 28-30.

Le, H., & Schmidt, F. L. (2006). Correcting for indirect range restriction in meta-analysis: testing a new meta-analytic procedure. *Psychological Methods, 11,* 416-438.

LeBreton, J. M., Burgess, J. R., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar?. *Organizational Research Methods, 6,* 80-128.

Lee, S., & Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research, 37,* 319-343.

Linn, R. L. (1983). Pearson selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement, 20,* 1-15.

Lönnqvist, J. E., Paunonen, S., Verkasalo, M., Leikas, S., Tuulio-Henriksson, A., & Lönnqvist, J. (2007). Personality characteristics of research volunteers. *European Journal of Personality, 21,* 1017-1030.

Lounsbury, J. W., Welsh, D. P., Gibson, L. W., & Sundstrom, E. (2005). Broad and narrow personality traits in relation to cognitive ability in adolescents. *Personality and Individual Differences, 38,* 1009-1019.

Lundström, S., Chang, Z., Råstam, M., Gillberg, C., Larsson, H., Anckarsäter, H., et al. (2012). Autism spectrum disorders and autistic-like traits: Similar etiology in the extreme end and the normal variation. *Archives of General Psychiatry, 69,* 46-52.

Lykken, D. T., Bouchard, T. J., McGue, M., & Tellegen, A. (1990). The Minnesota twin family registry: Some initial findings. *Acta Geneticae Medicae et Gemellologiae: Twin Research, 39,* 35-70.

Maenner, M.J., Rice, C.E., Arneson, C.L., Cunniff, C., Schieve, L. A., Carpenter, L. A., et al. (2014). Potential impact of DSM-5 Criteria on Autism Spectrum Disorder prevalence estimates. *JAMA Psychiatry, 71,* 292-300.

Madhyastha, T. M., Hunt, E., Deary, I. J., Gale, C. R., & Dykiert, D. (2009). Recruitment modeling applied to longitudinal studies of group differences in intelligence. *Intelligence, 37,* 422-427.

Mandy, W., Charman, T., Puura, K., & Skuse, D. (2014). Investigating the cross-cultural validity of DSM-5 autism spectrum disorder: Evidence from Finnish and UK samples. *Autism, 18,* 45-54.

Marcus, B., & Schütz, A. (2005). Who are the people reluctant to participate in research? Personality correlates of four different types of nonresponse as inferred from self-and observer ratings. *Journal of Personality, 73,* 959-984.

Maric, N., Myin-Germeys, I., Delespaul, P., de Graaf, R., Vollebergh, W., & Van Os, J. (2004). Is our concept of schizophrenia influenced by Berkson's bias?. *Social Psychiatry and Psychiatric Epidemiology, 39,* 600-605.

Martin, N. (2000). Gene-environment interaction and twin studies. In Spector, T., Sneider, H., & MacGregor, A. (Eds). *Advances in twin and sib-pair analysis.* Greenwich Medical Media: London, UK.

Martin, N. G., & Wilson, S. R. (1982). Bias in the estimation of heritability from truncated samples of twins. *Behavior Genetics, 12,* 467-472.

Marsh, H. W., Trautwein, U., Lüdtke, O., Baumert, J., & Köller, O. (2007). The big-fish-little-pond effect: Persistent negative effects of selective high schools on self-concept after graduation. *American Educational Research Journal, 44,* 631-669.

Mather, K., & Jinks, J. L. (1971). Biometrical genetics. *Biometrical Genetics (Ed. 2).* Chapman and Hall: London, UK.

Mazefsky, C.A., Goin-Kochel, R.P., Riley, B.P., & Maes, H.H. (2008). Genetic and environmental influences on symptom domains in twins and siblings with autism. *Research in Autism Spectrum Disorders, 2,* 320-331.

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review, 15,* 28-50.

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: a case for nonparametric item response theory modeling. *Psychological Methods, 9,* 354-368.

Meijer, R. R., & Egberink, I. J. (2012). Investigating Invariant Item Ordering in Personality and Clinical Scales Some Empirical Findings and a Discussion. *Educational and Psychological Measurement, 72,* 589-607.

Mein, G., Johal, S., Grant, R., Seale, C., Ashcroft, R., & Tinker, A. (2012). Predictors of two forms of attrition in a longitudinal health study involving ageing participants: An analysis based on the Whitehall II study. *BMC Medical Research Methodology*, *12,* 164.

Meriac, J.P., Hoffman, B.J., Woehr, D.J., & Fleisher, M.S. (2008). Further Evidence for the Validity of Assessment Center Dimensions: A Meta-Analysis of the Incremental Criterion-Related Validity of Dimension Ratings. *Journal of Applied Psychology*, *93*, 1042-1052.

Merikangas, K. R., He, J. P., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., ... & Swendsen, J. (2010). Lifetime prevalence of mental disorders in US adolescents: results from the National Comorbidity Survey Replication–Adolescent Supplement (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry*, *49*, 980-989.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156-166.

Miller, T. Q. (1994). High-risk studies are influenced by indirect range restriction. *Journal of Behavioral Medicine*, *17*, 567-588.

Molenaar, D., & Dolan, C. V. (2014). Testing systematic genotype by environment interactions using item level data. *Behavior Genetics*, *44*, 212-231.

Molenaar, D., van der Sluis, S., Boomsma, D. I., & Dolan, C. V. (2012). Detecting specific genotype by environment interactions using marginal maximum likelihood estimation in the classical twin design. *Behavior Genetics,42,* 483-499.

Morse, B. J., Johanson, G. A., & Griffeth, R. W. (2012). Using the graded response model to control spurious interactions in moderated multiple regression. *Applied Psychological Measurement, 36,* 122-146.

Moutafi, J., Furnham, A., & Crump, J. (2003). Demographic and personality predictors of intelligence: a study using the neo personality inventory and the Myers–Briggs type indicator. *European Journal of Personality, 17,* 79–94

Moutafi, J., Furnham, A., & Crump, J. (2006). What facets of openness and conscientiousness predict fluid intelligence scores. *Learning and Individual Differences.* 31-42.

Moutafi, J., Furnham, A., & Paltiel, L. (2004). Why is conscientiousness negatively correlated with intelligence?. *Personality and Individual Differences,37,* 1013-1022.

Murray, A.L., Allison, C., Smith, P., Booth, T., Baron-Cohen, S., Auyeung, B. Investigating diagnostic bias in autism spectrum conditions: An item response theory analysis of sex bias in the AQ-10. Submitted.

Murray, A. L., Booth, T., McKenzie, K., Kuenssberg, R., & O'Donnell, M. (2014). Are autistic traits measured equivalently in individuals with and without an autism spectrum disorder? An invariance analysis of the Autism Spectrum Quotient Short Form. *Journal of Autism and Developmental Disorders, 44,* 55-64.

Murray, A. L., Eisner, M., Ribeaud, D. In search of trans-diagnostic dimensional measures of

childhood and adolescent psychopathology: An analysis of the Social Behavior Questionnaire. *Submitted.*

Murray, A., McKenzie, K., & Murray, G. (2014). To what extent does g impact on conceptual, practical and social adaptive functioning in clinically referred children?. *Journal of Intellectual Disability Research, 58*, 777-785.

Murray, A. L., McKenzie, K. & Murray, K. R. (2015). An evaluation of the performance of the WISC-IV eight-subtest short form with children who may have an intellectual disability. *Journal of Intellectual & Developmental Disability.* In press.

Murray, A. L., & McKenzie, K. (2014). The accuracy of the Child and Adolescent Intellectual Disability Screening Questionnaire (CAIDS-Q) in classifying severity of impairment: a brief report. *Journal of Intellectual Disability Research*, 58, 1179-1184.

Muthén, B. O. (1990). Moments of the censored and truncated bivariate normal distribution. *British Journal of Mathematical and Statistical Psychology*, *43*, 131-143.

Muthén, B. O. (1989). Tobit factor analysis†. *British Journal of Mathematical and Statistical Psychology*, *42*, 241-250.

Muthén, L. K., & Muthén, B. O. (2010). Mplus User's Guide: Statistical Analysis with Latent Variables: User's Guide. Muthén & Muthén.

Neale M. C., Boker S. M., Xie G., Maes H. H. (2006) *Mx: statistical modeling, 7th edn.* VCU Department of Psychiatry, Richmond.

Neale, M. C., Eaves, L. J., Kendler, K. S., & Hewitt, J. K. (1989). Bias in correlations from selected samples of relatives: The effects of soft selection. *Behavior Genetics*, *19*, 163-169.

Nie, L., Chu, H., & Korostyshevskiy, V. R. (2008). Bias reduction for nonparametric correlation coefficients under the bivariate normal copula assumption with known detection limits. *Canadian Journal of Statistics, 36,* 427-442.

Nishiwaki, Y., Clark, H., Morton, S. M., & Leon, D. A. (2005). Early life factors, childhood cognition and postal questionnaire response rate in middle age: the Aberdeen Children of the 1950s study. *BMC Medical Research Methodology,5*, 16.

Nydick, S. W., (2014). catIrt: An R Package for Simulating IRT-Based Computerized Adaptive Tests. R package version 0.5-0. http://CRAN.R-project.org/package=catIrt

Ones, D. S., & Viswesvaran, C. (1996). Bandwidth–fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior, 17,* 609-626.

Osgood, D. W., McMorris, B. J., & Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance I: Item response theory scaling. *Journal of Quantitative Criminology, 18,* 267-296.

Pepermans, R., Vloeberghs, D. and Perkisas, B. (2003), High potential identification policies: An empirical study among Belgian companies. *Journal of Management Development, 22*, 660-78.

Paloutzian, R. F., & Ellison, C. W. (1982). Loneliness, spiritual well-being and the quality of life. In L. A. Peplau & D. Perlman (Eds.), *Loneliness: A sourcebook of current theory, research and therapy.* New York, NY: Wiley.

Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs.

Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 200, 1-66.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. *Assessment, 18,* 263-283.

Pluess, M., & Belsky, J. (2013). Vantage sensitivity: Individual differences in response to positive experiences. *Psychological Bulletin*, *139*, 901-916.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*, 322-338.

Purcell, S. (2002). Variance components models for gene–environment interaction in twin analysis. *Twin Research, 5*, 554-571.

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, ISBN 3-900051-07-0, URL http://www.R-project.org/

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German.*Journal of Research in Personality*, *41*, 203-212.

Rathouz, P. J., Van Hulle, C. A., Rodgers, J. L., Waldman, I. D., & Lahey, B. B. (2008). Specification, testing, and interpretation of gene-by-measured-environment interaction models in the presence of gene–environment correlation. *Behavior Genetics, 38*, 301-315.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27-48.

Reiss, D., Leve, L. D., & Neiderhiser, J. M. (2013). How genes and the social environment moderate each other. *American Journal of Public Health, 103*, S111-S121.

Rende, R., & Plomin, R. (1992). Diathesis-stress models of psychopathology: A quantitative genetic perspective. *Applied and Preventive Psychology, 1,* 177-182.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*, 145-154.

Rhemtulla, M., & Little, T. D. (2012). Planned missing data designs for research in cognitive development. *Journal of Cognition and Development,13*, 425-438.

Robinson, E.B., Koenen, K C., McCormick, M.C., Munir, K., Hallett, V., Happé, F., et al. (2012). A multivariate twin study of autistic traits in 12-year-olds: testing the fractionable autism triad hypothesis. *Behavior Genetics, 42*, 245-255.

Rutter, M. (2014). Addressing the issue of fractionation in autism spectrum disorder: A commentary on Brunsdon and Happé, Frazier et al., Hobson and Mandy et al. *Autism, 18,* 55-57.

Sackett, P. R., Laczo, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion interrater reliability: Implications for validation research. *Personnel Psychology, 55*, 807-825.

Sackett, P. R., Lievens, F., Berry, C. M., & Landers, R. N. (2007). A cautionary note on the effects of range restriction on predictor intercorrelations. *Journal of Applied Pssychology, 92*, 538-544.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: an expanded typology. *Journal of Applied Psychology, 85,* 112-118.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34,* 100.

Schmidt, F. L., & Hunter, J.E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529-540.

Schmidt, F. L., Oh, I. S., & Le, H. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology, 59,* 281-305.

Schmidt, F. L., Shaffer, J. A., & Oh, I. S. (2008). Increased accuracy for range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology, 61,* 827-868.

Schwabe, I., & van den Berg, S. M. (2014). Assessing genotype by environment interaction in case of heterogeneous measurement error. *Behavior Genetics, 44,* 394-406.

Shanahan, M. J., & Hofer, S. M. (2005). Social context in gene–environment interactions: Retrospect and prospect. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 60,* 65-76.

Silventoinen, K., Hasselbalch, A. L., Lallukka, T., Bogl, L., Pietiläinen, K. H., Heitmann, B. L., … & Kaprio, J. (2009). Modification effects of physical activity and protein intake on heritability of body size and composition. *The American journal of Clinical Nutrition, 90,* 1096-1103.

South, S. C., Krueger, R. F., Johnson, W., & Iacono, W. G. (2008). Adolescent personality moderates genetic and environmental influences on relationships with parents. *Journal of Personality and Social Psychology, 94*, 899-912.

South, S. C., & Krueger, R. F. (2011). Genetic and environmental influences on internalizing psychopathology vary as a function of economic status. *Psychological Medicine*, *41*, 107-117.

South, S. C., & Krueger, R. F. (2014). Genetic strategies for probing conscientiousness and its relationship to aging. *Developmental Psychology,50,* 1362-1376.

Stewart, M. E., Allison, C., Baron-Cohen, S., & Watson, R. (2015). Investigating the Structure of the Autism-Spectrum Quotient Using Mokken Scaling. *Psychological Assessment, 37,* 596-604.

Stone, E. F., & Hollenbeck, J. R. (1989). Clarifying some controversial issues surrounding statistical procedures for detecting moderator variables: Empirical evidence and related matters. *Journal of Applied Psychology, 74*, 3-10.

Stoolmiller, M. (1998). Correcting estimates of shared environmental variance for range restriction in adoption studies using a truncated multivariate normal model. *Behavior Genetics*, *28*, 429-441.

Studer, J., Baggio, S., Mohler-Kuo, M., Dermota, P., Gaume, J., Bertholet, N., ... & Gmel, G. (2013). Examining non-response bias in substance use research—are late respondents proxies for non-respondents?. *Drug and Alcohol Dependence*, *132*, 316-323.

Tabery, J. (2008). RA Fisher, Lancelot Hogben, and the origin (s) of genotype–environment interaction. *Journal of the History of Biology, 41, 717*-761.

Takagishi, H., Takahashi, T., Yamagishi, T., Shinada, M., Inukai, K., Tanida, S., et al. (2010). Salivary testosterone levels and autism-spectrum quotient in adults. *Neuroendocrinology Letters, 31,* 101-105.

Taylor, A. (2004). The consequences of selective participation on behavioral-genetic findings: Evidence from simulated and real data. *Twin Research, 7,* 485-504.

Tellegen, A., & Waller, N. G. (2008). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire.*The SAGE handbook of personality theory and assessment, 2,* 261-292.

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment, 18,* 291-307.

Timberlake, D. S., Rhee, S. H., Haberstick, B. C., Hopfer, C., Ehringer, M., Lessem, J. M., ... & Hewitt, J. K. (2006). The moderating effects of religiosity on the genetic and environmental determinants of smoking initiation. *Nicotine & Tobacco Research, 8,* 123-133.

Tucker-Drob, E. M., Harden, K. P., & Turkheimer, E. (2009). Combining nonlinear biometric and psychometric models of cognitive abilities. *Behavior Genetics, 39,* 461-471.

Tuvblad, C., Grann, M., & Lichtenstein, P. (2006). Heritability for adolescent antisocial behaviour differs with socioeconomic status: gene–environment interaction. *Journal of Child Psychology and Psychiatry, 47,* 734-743.

Uttl, B. (2005). Measurement of individual differences lessons from memory assessment in research and clinical practice. *Psychological Science, 16,* 460-467.

van den Berg, S. M., Glas, C. A., & Boomsma, D. I. (2007). Variance decomposition using an IRT measurement model. *Behavior Genetics, 37*, 604-616.

van den Oord, E. J., Pickles, A., & Waldman, I. D. (2003). Normal variation and abnormality: an empirical study of the liability distributions underlying depression and delinquency. *Journal of Child Psychology and Psychiatry,44*, 180-192.

van der Sluis, S., Dolan, C. V., Neale, M. C., Boomsma, D. I., & Posthuma, D. (2006). Detecting genotype–environment interaction in monozygotic twin data: comparing the Jinks and Fulker test and a new test based on marginal maximum likelihood estimation. *Twin Research and Human Genetics, 9*, 377-392.

van der Sluis, S., Posthuma, D., & Dolan, C. V. (2012). A note on false positives and power in G× E modelling of twin data. *Behavior Genetics, 42*, 170-186.

Van Goor, H., Jansma, F., & Veenstra, R. (2005). Differences in undercoverage and nonresponse between city neighbourhoods in a telephone survey. *Psychological Reports, 96*, 867-878.

van Hulle, C. A., Lahey, B. B., & Rathouz, P. J. (2013). Operating characteristics of alternative statistical methods for detecting gene-by-measured environment interaction in the presence of gene–environment correlation in twin and sibling studies. *Behavior Genetics, 43*, 71-84.

Verdoux, H., & van Os, J. (2002). Psychotic symptoms in non-clinical populations and the continuum of psychosis. *Schizophrenia Research, 54*, 59-65.

Vink, J. M., Willemsen, G., Stubbe, J. H., Middeldorp, C. M., Ligthart, R. S., Baas, K. D., ... & Boomsma, D. I. (2004). Estimating non-response bias in family studies: application to mental health and lifestyle. *European Journal of Epidemiology, 19*, 623-630.

Volken, T. (2013). Second-stage non-response in the Swiss health survey: determinants and bias in outcomes. *BMC Public Health, 13,* 167.

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. *Multivariate Behavioral Research,43,* 476-496.

Wechsler, D. (1974). *WAIS-R manual: Wechsler adult intelligence scale-revised*. Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children (4th ed.).* London, UK: Psychological Corporation.

Weiss, A., & Costa, P. T. (2014). Re:"Personality and All-Cause Mortality: Individual-Participant Meta-Analysis of 3,947 Deaths in 76,150 Adults".*American Journal of Epidemiology, 179,* 791-792.

Wheelwright, S., Baron-Cohen, S., Goldenfeld, N., Delaney, J., Fine, D., Smith, R., ... & Wakabayashi, A. (2006). Predicting autism spectrum quotient (AQ) from the systemizing quotient-revised (SQ-R) and empathy quotient (EQ). *Brain Research, 1079,* 47-56.

Wheelwright, S., Auyeung, B., Allison, C., & Baron-Cohen, S. (2010). Research defining the broader, medium and narrow autism phenotype among parents using the Autism Spectrum Quotient (AQ). *Molecular Autism, 1*.

Whitaker, S., & Gordon, S. (2012). Floor effects on the WISC-IV. *International Journal of Developmental Disabilities, 58,* 111-119.

Williams, D.M., & Bowler, D.M. (2014). Autism spectrum disorder: Fractionable or coherent? *Autism, 18,* 2-5.55.

Wood, P., & Englert, P. (2009). Intelligence compensation theory: A critical examination of the negative relationship between conscientiousness and fluid and crystallised intelligence. *The Australian and New Zealand Journal of Organisational Psychology*, *2*, 19-29.

Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A., Dudbridge, F., & Middeldorp, C. M. (2014). Research Review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*,*55*, 1068-1087.

Wrulich, M., Brunner, M., Stadler, G., Schalke, D., Keller, U., & Martin, R. (2014). Forty years on: Childhood intelligence predicts health in middle adulthood. *Health Psychology*, *33*, 292-296.

Yang, H., Sackett, P. R., & Nho, Y. (2004). Developing a procedure to correct for range restriction that involves both institutional selection and applicants' rejection of job offers. *Organizational Research Methods*, *7*, 442-455.

Zaroff, C. M., & Uhm, S. Y. (2012). Prevalence of autism spectrum disorders and influence of country of measurement and ethnicity. *Social Psychiatry and Psychiatric Epidemiology*, *47*, 395-398

Zheng, H. & Rathouz, P. (2013). GxM: Maximum Likelihood Estimation for Gene-by-Measured Environment Interaction Models. R package version 1.0. http://CRAN.Rproject.org/package=GxM.

Appendix: R code for GxM simulations

```
#data generation

make.irt.data<-function(filenames){

  N=1000  ##sample size
  am=0.3^0.5   #for moderator
  cm=0.2^0.5
  em=0.5^0.5
  ac=0.3^0.5   #common to moderator and phenotype
  cc=0.1^0.5
  ec=0.1^0.5
  au=0.2^0.5   #unique to phenotype
  cu=0.1^0.5
  eu=0.2^0.5
  alpha_c=0   #moderation of common A
  gamma_c=0   #moderation of common C
  epsilon_c=0  #moderation of common E
  alpha_u=-0.15 #moderation of unique A
  gamma_u=0.20    #moderation of unique C
  epsilon_u=0.08  #moderation of unique E


  ####In this section variance-covariance matrix the LVs in the GxM is
  ####defined for the MZ twins. A,C and E refer to source of variance
  ####c and u refer to common and unique
  ### 1 and 2 refer to twin 1 and twin 2


  MZ=matrix(c(
    1,0,0,0,0,0,0,0,0,0,0,0,        #Ac1
    0,1,0,0,0,0,0,0,0,0,0,0,        #Cc1
    0,0,1,0,0,0,0,0,0,0,0,0,        #Ec1
    0,0,0,1,0,0,0,0,0,0,0,0,        #Au1
    0,0,0,0,1,0,0,0,0,0,0,0,        #Cu1
    0,0,0,0,0,1,0,0,0,0,0,0,        #Eu1
    1,0,0,0,0,0,1,0,0,0,0,0,        #Ac2
    0,1,0,0,0,0,0,1,0,0,0,0,        #Cc2
    0,0,0,0,0,0,0,0,1,0,0,0,        #Ec2
    0,0,0,1,0,0,0,0,0,1,0,0,        #Au2
    0,0,0,0,1,0,0,0,0,0,1,0,        #Cu2
    0,0,0,0,0,0,0,0,0,0,0,1),12,12,byrow=T) #Eu2

  MZ=MZ+t(MZ)-diag(diag(MZ))
```

```
  dimnames(MZ)[[1]]<-
c("Ac1","Cc1","Ec1","Au1","Cu1","Eu1","Ac2","Cc2","Ec2","Au2","Cu2","Eu2")
  dimnames(MZ)[[2]]<-
c("Ac1","Cc1","Ec1","Au1","Cu1","Eu1","Ac2","Cc2","Ec2","Au2","Cu2","Eu2")

  ##generate latent variable scores by sampling from MVN distribution
  MZ_ACE=mvrnorm(N,c(0,0,0,0,0,0,0,0,0,0,0,0),MZ,empirical=T)
  dimnames(MZ_ACE)[[2]]<-
c("Ac1","Cc1","Ec1","Au1","Cu1","Eu1","Ac2","Cc2","Ec2","Au2","Cu2","Eu2")

  ##model for moderator
  ##this is from eq. 1 in the manuscript
  MZ_ACE<-as.data.frame(MZ_ACE)
  attach(MZ_ACE)
  M1_MZ<-Ac1*am + Cc1*cm + Ec1*em
  M2_MZ<-Ac2*am + Cc2*cm + Ec2*em

  ##M_MZ is the matrix of moderator scores for MZ twin1 and twin2
  M_MZ<-cbind(M1_MZ,M2_MZ)


  ##model for phenotype
  P1_MZ<-
ac*Ac1+cc*Cc1+ec*Ec1+(au+alpha_u*M1_MZ)*Au1+(cu+gamma_u*M1_MZ)*Cu1+(eu+
epsilon_u*M1_MZ)*Eu1
  P2_MZ<-
ac*Ac2+cc*Cc2+ec*Ec2+(au+alpha_u*M2_MZ)*Au2+(cu+gamma_u*M2_MZ)*Cu2+(eu+
epsilon_u*M2_MZ)*Eu2


  detach(MZ_ACE)
  P_MZ<-cbind(P1_MZ,P2_MZ)
  MZdata<-cbind(M_MZ,P_MZ)

  ####DZs
  DZ=matrix(c(
    1,0,0,0,0,0,0,0,0,0,0,0,        #Ac
    0,1,0,0,0,0,0,0,0,0,0,0,        #Cc
    0,0,1,0,0,0,0,0,0,0,0,0,        #Ec
    0,0,0,1,0,0,0,0,0,0,0,0,        #Au
    0,0,0,0,1,0,0,0,0,0,0,0,        #Cu
    0,0,0,0,0,1,0,0,0,0,0,0,        #Eu
    .5,0,0,0,0,0,1,0,0,0,0,0,        #Ac
    0,1,0,0,0,0,0,1,0,0,0,0,        #Cc
    0,0,0,0,0,0,0,0,1,0,0,0,        #Ec
    0,0,0,.5,0,0,0,0,0,1,0,0,        #Au
    0,0,0,0,1,0,0,0,0,0,1,0,        #Cu
    0,0,0,0,0,0,0,0,0,0,0,1),12,12,T) #Eu

  DZ=DZ+t(DZ)-diag(diag(DZ))
```

```
  dimnames(DZ)[[1]]<-
c("Ac1","Cc1","Ec1","Au1","Cu1","Eu1","Ac2","Cc2","Ec2","Au2","Cu2","Eu2")
  dimnames(DZ)[[2]]<-
c("Ac1","Cc1","Ec1","Au1","Cu1","Eu1","Ac2","Cc2","Ec2","Au2","Cu2","Eu2")

  ##generate latent variable scores
  ACE_DZ=mvrnorm(N,c(0,0,0,0,0,0,0,0,0,0,0,0),DZ,empirical=T)
  dimnames(ACE_DZ)[[2]]<-
c("Ac1","Cc1","Ec1","Au1","Cu1","Eu1","Ac2","Cc2","Ec2","Au2","Cu2","Eu2")

  ##model for moderator
  ACE_DZ<-as.data.frame(ACE_DZ)
  attach(ACE_DZ)
  M1_DZ <- Ac1*am + Cc1*cm + Ec1*em
  M2_DZ <- Ac2*am + Cc2*cm + Ec2*em

  M_DZ=cbind(M1_DZ,M2_DZ)

  ##model for phenotype
  P1_DZ<-
(ac+alpha_c*M1_DZ)*Ac1+(cc+gamma_c*M1_DZ)*Cc1+(ec+epsilon_c*M1_DZ)*Ec1+(a
u+alpha_u*M1_DZ)*Au1+(cu+gamma_u*M1_DZ)*Cu1+(eu+epsilon_u*M1_DZ)*Eu1
  P2_DZ<-
(ac+alpha_c*M2_DZ)*Ac2+(cc+gamma_c*M2_DZ)*Cc2+(ec+epsilon_c*M2_DZ)*Ec2+(a
u+alpha_u*M2_DZ)*Au2+(cu+gamma_u*M2_DZ)*Cu2+(eu+epsilon_u*M2_DZ)*Eu2

  detach(ACE_DZ)
  P_DZ<-cbind (P1_DZ,P2_DZ)
  DZdata<-cbind(M_DZ,P_DZ)


  ###data based on P and M
  MZDZ<-as.data.frame(rbind(MZdata,DZdata))
  MZDZ<-rename(MZDZ, c('M1_MZ'='M1','M2_MZ'='M2','P1_MZ'='P1','P2_MZ'='P2'))



  ######generate data according to GRM
  params<-cbind(a=c(2.44,1.15,1.93,1.96,2.13,1.09,0.67,1.13,0.87,0.99,1.01,1.63,
                    1.75,0.80,1.91,0.55,1.06,1.88,0.90,1.94),
                b1=c(-0.27,-0.21,-0.11,-0.36,0.34,-0.15,0.34,0.23,0.43,
                     0.04,0.10,0.01,0.37, 0.13,0.00,0.50,-0.24,-0.40,-0.11,-0.24),
                b2=c(0.84,1.46,1.50,1.29,1.16,1.34,0.99,0.68,0.98,1.22,0.93,0.67,1.49,0.89,
                     1.29,0.76,1.02,0.80,1.27,0.65),
                b3=c(2.23,2.01,2.38,2.07,2.07,2.00,2.34,2.33,2.22,2.39,2.27,2.20,2.42,2.29,
                     2.09,2.32,2.07,2.09,2.27,2.17),
                b4=c(2.74,2.73,2.82,2.65,2.73,2.78,2.64,2.62,2.83,2.73,2.63,
                     2.75,2.67,2.92,2.96,2.81,2.74,2.86,2.73,2.73))
```

```
  P1<-MZDZ$P1
  P1<-as.numeric(P1)
  P1_items<-simIrt(theta=P1, params=params, mod='grm')
  P1_i<-P1_items$resp



  P2<-MZDZ$P2
  P2<-as.numeric(P2)
  P2_items<-simIrt(theta=P2, params=params, mod='grm')
  P2_i<-P2_items$resp


  ####fit IRT model####
  P1_grm<-mirt(data=P1_i, model=1, itemtype='graded')
  P1_fs<-fscores(P1_grm, method='EAP', full.scores=T, scores.only=T)#,
response.pattern=P1_i)
  P1_fs_z<-scale(P1_fs)


  P2_grm<-mirt(data=P2_i, model=1, itemtype='graded')
  P2_fs<-fscores(P2_grm, method='EAP', full.scores=T, scores.only=T)#,
response.pattern=P1_i)
  P2_fs_z<-scale(P2_fs)


  MZDZ$zyg<-c(rep(1,N),rep(2,N))


  MZDZ_IRT<-cbind(MZDZ$zyg, scale(P1_fs_z), scale(P2_fs_z), MZDZ$M1, MZDZ$M2,
MZDZ$M1, MZDZ$M2)

  write.table(MZDZ_IRT, file=filenames, col.names=F, row.names=F, sep='  ')

}
filenames<-as.matrix(c(paste('D:/GxM polytomous
1000/UE_1_IRT_poly_1000',c(1:100),'.dat',sep=' ')))
apply(filenames, 1, make.irt.data)
```